BLIND DECONVOLUTION AND PHASE RETRIEVAL

A thesis presented for the degree of
Doctor of Philosophy
in Electrical & Electronic Engineering
in the
University of Canterbury,
Christchurch, New Zealand.

by
Richard Lane
February 1988

Errata

The following errors should be corrected in the text:
page 62: (Levi and Stark 1982) should be (Sezan and Stark 1982).
page 73: The last paragraph of section 4.3 following (4.20) should read:
which correspond to the images shown in Fig 4.10. Since Fig 4.10c is equivalent to 4.9a rotated 180 degrees around the coordinate origin they both relate to the same image-form. By similar reasoning 4.10a and 4.10b also have the same image-form, but this image-form is clearly different to the image-form shown in Figs 4.9a and 4.10c. In general, if the true image is the convolution of N components then there are 2^N image-forms (Lane et al 1987).
page 74: Figs 4.9a and 4.9b should be interchanged.
page 82: Fig 4.17 should be rotated clockwise by 90 degrees.
page 83: Figs 4.18a and 4.18b should be rotated clockwise by 270 degrees.
page 93: (Fienup and Dainty 1986) should be (Dainty and Fienup 1987).
page 122: (Tan and Bates 1985) should be (Bates and Tan 1985).
page 140: (Oppenheim et al. 1980) should be (Oppenheim et al. 1982).

Abstract

Theoretical and practical aspects of identifying and deconvolving a convolution in more than one-dimension are presented. In contrast to conventional techniques which require knowledge of the blurring function, this thesis describes techniques for "blind" deconvolution. The techniques introduced differ from previous work in the field of blind deconvolution because they do not require an ensemble of similarly blurred images, i.e. they can be effectively employed upon a single convolution.

The first method for blind deconvolution introduced relies on the analytic properties of the Fourier spectrum of a compact image. Rather than deal with continuous images, a discrete approximation is employed. It is argued, however, that approximation of the Fourier spectrum by a finite order polynomial model is a logical response to the practical constraints posed by limited amounts of noisy data.

Since the convolution of two images is equivalent to a multiplication of their Fourier spectra, deconvolution is consequently equivalent to factorisation of their Fourier spectra. In one dimension it is always possible to factorise a polynomial, even when it is of infinite order. These factors correspond to isolated points, in the complex plane into which the Fourier spectra are analytically continued, where the spectra are zero. Since these points are distinct there are a large number of factors and hence there is usually a large number of ways of deconvolving a one-dimensional image.

By contrast the analytically continued Fourier spectrum of a two-dimensional image exists in a four-dimensional space and is zero on a two-dimensional analytic surface, here called a zero-sheet. Because of the analytic nature of the zero-sheet it is not possible, in general, to factorise a two-dimensional spectrum or equivalently partition its zero-sheet into separate analytic surfaces. The major exception is when the true image is a convolution in which case the zero-sheet is, in fact, the union of the zero-sheets of the components of the convolution. As a result the zero-sheet of a convolution can be partitioned into two zero-sheets which can be used to recover, to within a complex constant, the components of the convolution. The addition of noise is shown to link the zero-sheets of the components of the convolution. Consequently it is no longer possible to partition the zero-sheet without isolating and correcting these "bridges" between the zero-sheets of the components.

The Fourier phase problem forms a special subclass of the blind deconvolution problem, one in which the true image and the blurring function are conjugate mirror images of each other. The data in the Fourier phase problem comprises the oversampled magnitude of the Fourier transform of the true image. Consequently, it is necessary to reconstruct the Fourier phase before an estimate of the true image can be formed. It is shown that a solution exists and the accuracy of the solution can be empirically related to the amount of noise present in the Fourier magnitude data.

It is shown that a unique solution to the Fourier phase problem in more than one

dimension exists except when the spectrum is the Fourier transform of a convolution. In this case, the number of solutions to the Fourier phase problem is related to the number of component images which have been convolved to produce the convolution.

The second technique for deconvolution introduced in this thesis uses these multiple solutions to the Fourier phase problem to recover information about the phase of the spectra of the components of the convolution. The Fourier phase is, however, only recovered modulo π. The problems encountered in the modified magnitude problem, as it is called in this thesis, are analysed and techniques for overcoming these difficulties are described.

A final result presented herein is an extension to an existing technique for blind deconvolution of ensembles of two-dimensional speckle images. It is shown that comparing the zero-sheets of the speckle spectra leads to a useful new approach to speckle imaging.

Acknowledgements

I am indebted to many people for helping with the research contained in this thesis. I would like to thank them all, especially my supervisor Richard Bates. I would also like to thank all those who have made my stay in Christchurch both enjoyable and memorable.

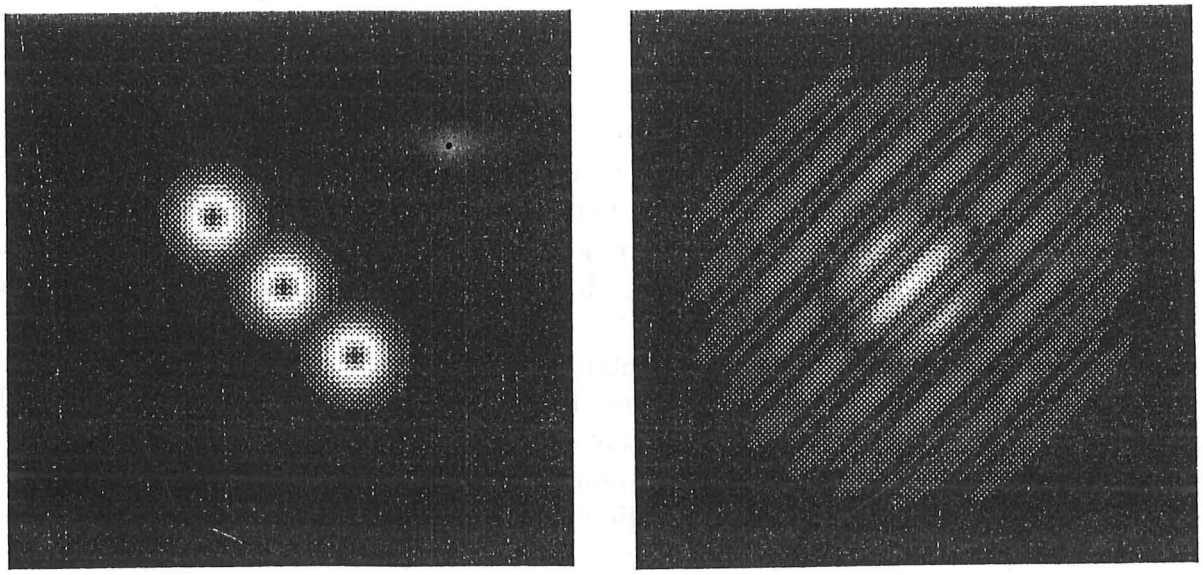
Preface

A common problem in engineering is that complete measurements of a given quantity or object are often difficult, expensive or perhaps even impossible to make directly. Two extreme examples of these difficulties are found in astronomy and crystallography. Direct measurements in the former are constrained by vast distances and the intervening atmosphere, whilst in the latter they are inhibited by the minute distances involved. In both instances only partial information is available on the true nature of the object. In radio astronomy, for example, often only the magnitude of the incident radiation is directly measurable. This bears no visual resemblance to the actual astronomical object, but is related to what is known as the object's Fourier spectrum through a unique invertible transform. Although there is a unique relationship between an object and its Fourier spectrum, discussed in detail in chapter 1, an estimate of the phase of the incident radiation is required before it is possible to reconstruct the object of interest.

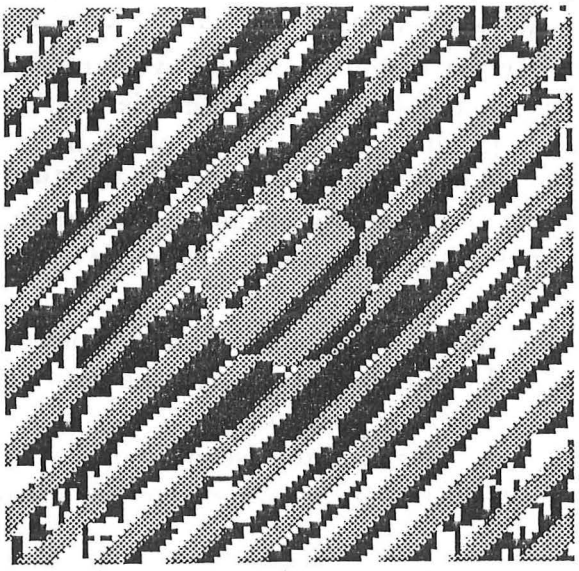
One of the main applications of the work reported in this thesis is the problem of inferring the Fourier phase from measurements of the Fourier magnitude. The inverse Fourier transform can then be invoked to recover the original object. The algorithms in this thesis require the Fourier magnitude to be oversampled, which is nearly always possible in applications where the Fourier transform is continuous, such as astronomy and wavefront sensing. Reconstructing the Fourier phase from the oversampled Fourier magnitude constitutes the Fourier phase problem (Bates 1982b). In crystallography, however, the Fourier transform exists only at discrete points in Fourier space called the Nyquist frequencies. The crystallographic phase problem is thus only discussed in passing.

Figure 1 shows an example of an object and its Fourier transform. The Fourier magnitude consists of parallel bands between which the Fourier magnitude falls to zero. By comparing the Fourier magnitude with the phase it can be seen that regions where the Fourier magnitude is zero correspond to discontinuities in the Fourier phase. In this simple case, at least, it is clear that there is a strong relationship between the Fourier magnitude and the Fourier phase.

The relationship between the one-dimensional Fourier intensity and Fourier phase has been well understood for many years, having previously found application in the fields of control, filter and antenna design (Nagrath and Gopal 1975; Oppenheim and Schafer 1975; Bates 1969; Taylor and Whinnery 1951). This work had shown that, although it is possible to retrieve a one-dimensional Fourier phase from a given Fourier magnitude, there are in fact many mathematically acceptable solutions. The technique for recovering these solutions involves expressing the Fourier magnitude as a polynomial. A one-dimensional polynomial can be factored in a number of ways, each of which gives an alternative phase distribution consistent with the given Fourier magnitude (in a manner described in chapter 3).



(a) (b)



(c)

Figure 1: An example of a positive object and its Fourier transform. (a) the positive object quantised into 32 grey levels from 0 (black) to a normalised magnitude of 1 (white). (b) The magnitude of the Fourier transform of (a) quantised as above. (c) The phase of the Fourier transform of (a) quantised from $-\pi$ (black) to π (white).

In more than one dimension the Fourier phase problem appeared more obscure to begin with because the factoring of two-dimensional polynomials is a difficult problem. Two-dimensional recursive filtering is a field where the inability to factorise multi-dimensional polynomials poses a significant difficulty in the practical implementation of recursive digital filters (Huang et al 1971; Rabiner and Gold 1975, chapter 7). Bruck and Sodin (1979) were the first to note that the absence of a fundamental theorem of algebra in higher dimensions could be used to advantage in phase retrieval (or the phase problem) where the lack of factors would avoid the inherent ambiguities of the one-dimensional case. The significance of Bruck and Sodin's work was that it showed, theoretically at least, that unique multi-dimensional phase retrieval was worth pursuing. This confirmed the early hypothesis of Professor R.H.T. Bates and Peter J. Napier at the University of Canterbury who in the early 1970's (Napier and Bates 1974) suggested that there may be a unique solution to the two-dimensional phase retrieval problem. Their argument used the projection theorem, commonly applied in computed tomography (Garden 1984), to reduce the two-dimensional problem to a number of related one-dimensional problems. Although this work showed that the problem was not hopeless it did not produce an algorithm. W. Richard Fright used a different form of projection to argue for the unique relationship between the Fourier magnitude and the Fourier phase (Fright 1984, §3.5) in two dimensions. The method relied on calculating the difference in phase between two arbitrary points in the two-dimensional Fourier transform along several different paths. As this phase difference should be independent of path, inconsistent solutions can be eliminated.

A phase recovery algorithm employing exact phase closure was first implemented by H.V. Deighton, M.S. Scivier and Michael A. Fiddy at Imperial College London (Deighton et al 1985), and a similar algorithm was programmed by the author. The difficulty with these methods is that the one-dimensional projections are all individually subject to a large ambiguity and the amount of computer time required to find the consistent solution rises exponentially with image size. An earlier attempt at a direct algorithm produced at the University of Canterbury, was the CPE algorithm (Bates 1982b; Fright and Bates 1982). CPE used phase closure in a similar manner to Fiddy's technique to provide an estimate of the two-dimensional phase. The phase differences were estimated in a crude, but for simple images, effective manner. This algorithm solved the computational difficulties but at a price in accuracy. Further developments and extensions of CPE were provided by Daniel Mnyama and Michael C. Won, also of the University of Canterbury. For intricate images CPE has found use chiefly as a starting point to iterative methods of solution (Won et al 1985).

This thesis introduces a new direct solution of the Fourier phase problem (Lane and Bates 1987a; Lane et al 1987). The technique, discussed in chapter 4, relies on a new method of determining the factors of a multi-dimensional polynomial. The method can, by analysing where the Fourier transform is zero, determine how a given multi-dimensional polynomial can be factored.

Before leaving the subject of direct phase retrieval, other major contributors to the theoretical relationship between the Fourier magnitude and phase should be noted. These include Jorge L.C. Sanz at the IBM Research Laboratories, M. Hayes, J.S. Lim and A.V. Oppenheim at the Massachussets Institute of Technology whose work is discussed in more detail in the body of this thesis. Their algebraic descriptions complement the geometrical arguments and algorithms which comprise some of the original work presented

in this thesis.

Whilst the search continued for direct algorithms, iterative methods were developed for phase recovery. The initial methods were simple adaptions of existing algorithms and prone to stagnation when far from the true solution (Oppenheim and Lim 1981). This stagnation is a consequence of algorithmic ill-conditioning and had effectively precluded obtaining useful results before James R. Fienup successfully modified the existing algorithms to produce a method for phase retrieval for positive images(Fienup 1982). The causes of stagnation and Fienup's technique for improving convergence are discussed in chapter 5.

It was thought that Fienup's approach was limited to applications where the object was entirely real and positive (such as an intensity distribution) or of a special shape (Fienup 1987). In chapter 5 a new strategy in applying Fienup's existing algorithms is discussed (Lane 1987). This strategy extends the applicability of Fienup's methods to phase recovery for complex objects.

The major emphasis of this thesis is the extension of the above-mentioned phase retrieval work to the broader context of deconvolution. Convolution can be modelled as a process of multiplying the Fourier transforms of the images. Deconvolution therefore requires the factoring of general complex-valued Fourier spectra and is thus a natural generalisation of the factorisation of the Fourier magnitude required for phase retrieval. This provides the basis of the first of the new approaches to deconvolution introduced in this thesis and is discussed in chapter 4.

The second new approach to deconvolution is described in chapter 6. This algorithm arose from the observation that convolutions do not have a unique relationship between their Fourier magnitude and phase. There are, however, only a finite number of solutions, all of which can be obtained by iterative techniques. It proved possible to combine these ambiguous solutions to provide an estimate of the Fourier phase of the components of the convolutions. As the Fourier phase appeared to dominate the magnitude in determining the overall structure of the image, it had been previously thought that magnitude retrieval from phase was a relatively simple procedure (Bruck and Sodin 1983; Oppenheim and Lim 1981; Fright 1984). Although this is to a large extent true there are a number of significant practical difficulties associated with magnitude retrieval which are discussed in chapter 6.

This preface concludes with a brief description of the contents of the chapters herein. The mathematical basis of this work is to be found in chapter 1. This initial chapter is not intended to deter the less mathematically inclined reader, but to introduce a consistent notation which is used throughout this thesis. The chapter starts with a general discussion of the role of models, where it is emphasised that there are important differences between physical reality, mathematical models and the representation of models on a digital computer. As the Fourier transform is fundamental to the algorithms developed in this thesis, chapter 1 follows it from the idealised mathematical equations, through necessary practical assumptions to a discussion of some of the difficulties encountered in employing the Fourier transform on a digital computer.

Chapter 2 introduces the practical situations where the techniques of chapter 1 can be applied. The mathematical description of defocus and motion blur in photography is presented. This provides an introduction to the more complicated distortion caused by the earth's turbulent atmosphere. The phase and the magnitude problem are both introduced and discussed in the context of relevant practical situations.

In chapter 3, entire functions are introduced. They form the mathematical basis of the new techniques which are described in this thesis. Entire functions form a powerful class of models which can be thought of as generalised polynomials. The Fourier transform has been analysed in terms of entire functions by a number of researchers (Ross et al. 1978; Nakajima and Asakura 1983a; Burge et al. 1976). The discrete Fourier transform which provides the computational basis for the models used in this thesis, is an example of a particular class of entire function.

Chapter 4 describes the first new method of deconvolution. Two-dimensional entire functions are shown to be defined by their zero-sheets (regions where the modelling functions are zero). The zero-sheet of a convolution is shown to be the union of the zero-sheets of the convolution's components (of which there are usually two, one corresponding to the object of interest and the other to the distortion). Deconvolution can thus be effected by separation of these zero-sheets, each of which can be related to a component of a convolution. The effects of noise are considered and simple geometrical arguments are presented to complement the algebraic results of Hayes and Sanz on the mathematical properties of multi-dimensional polynomials.

Chapter 5 is concerned with the recovery of images from their Fourier magnitudes using the iterative Fienup algorithms. These algorithms are extended shown to be capable of reconstructing both bipolar and complex images. Complex images had previously been thought to be recoverable from their Fourier magnitudes only when very strict requirements on their support were applicable (Fienup 1987). Results are presented to show that the Fienup algorithms are in fact capable of achieving phase recovery for complex images, provided an alternative iteration strategy is employed.

Chapter 6 combines the iterative work on phase recovery with the mathematical properties of convolutions to form an alternative algorithm for deconvolution. Previous work (Fright 1984; Fiddy et al 1983; Van Toorn and Ferweda 1977) has described how the Fourier phase cannot be uniquely inferred from the Fourier magnitude of a convolution. The number of possible solutions is given by the number of components in the convolution. Thus the convolution of two images has two distinct phase distributions compatible with the convolution's magnitude, both of which can usually be recovered by using the Fienup algorithms. This chapter shows that the multiple phase distributions compatible with a convolution's Fourier magnitude, can be used as the basis of an algorithm for deconvolution.

The thesis concludes with recommendations for further research in chapter 7. Papers and presentations prepared during the course of this thesis are listed below in order of preparation.

1. Lane R. G. , Fright W. R. and Bates R. H. T. , "Direct phase retrieval", *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume ASSP-35, April 1987, pages 520–526.

2. Lane R. G. and Bates R. H. T. , "Automatic multi-dimensional deconvolution", *Journal Optical Society of America A*, Volume 3, January 1987, pages 180–184.

3. Lane R. G. , "Recovery of complex images from Fourier magnitude", *Optics Communications*, volume 63, 1987, pages 6–10.

4. Lane R. G. and Bates R. H. T. ", "Relevance for blind deconvolution of recovering Fourier magnitude", *Optics Communications*, volume 63, 1987, pages 11–14.

5. Bates R. H. T. and Lane R. G. , "Automatic deconvolution and phase retrieval", *Proceedings of the Joint Workshop on High-Resolution Imaging from the Ground Using Interferometric Techniques*, Oracle, Arizona, January 12 - 15, 1987, pages 71–73.

6. Bates R. H. T. and Lane R. G. , "Deblurring should now be automatic" 6th Pfefferkorn Conference on Image and Signal Processing in Electron Microscopy, 27 April to 2 May 1987, Niagara Falls, Ontario, Canada.

7. Bates R. H. T. and Lane R. G. , "Automatic deconvolution and phase retrieval", *Proceedings SPIE, vol 828, Optical and Optoelectronic Applied Science and Engineering* San Diego, August 1987.

Glossary

The notation employed in this thesis is used to describe three K-dimensional spaces called image-space, Fourier-space and Z-transform space. Arbitrary points in these spaces are described by the position vectors $\vec{x}$, $\vec{u}$, $\vec{\zeta}$, respectively. In two-dimensional coordinates these position vectors may be specialised to (x,y), (u,v) and (ζ,γ) respectively. The quantities existing in these spaces are respectively called images, Fourier-spectra and Z-spectra. Arbitrarily chosen Cartesian coordinates are set up in each space with x_k, u_k and ζ_k being the respective k^{th} coordinates.

The following conventions are adhered to in this thesis:

1. Symbols representing vector quantities are indicated by a superscript arrow, e.g. $(\vec{x})$.

2. Image quantities are defined by lower case roman symbols, e.g. $f(x,y)$.

3. Fourier transforms are written in upper case roman symbols, e.g. $F(u,v)$.

4. Z-transforms are indicated by script upper case letters, e.g. $\mathcal{F}(\zeta,\gamma)$.

5. An estimated quantity is adorned by a superscript caret, e.g. $\hat{f}(x,y)$ indicates an estimate of $f(x,y)$.

The following symbols are used:

$\longleftrightarrow$	a Fourier transform pair
$\overset{z}{\longleftrightarrow}$	a Z-transform pair
$\int(K)\int$	K-dimensional integral
$d\sigma(\ )$	K-dimensional volume element
$f^*(x,y)$	complex conjugate of $f(x,y)$
$\odot$	convolution
$\cup$	union of sets
$\cap$	intersection of sets
$< >$	ensemble average
$\{\ \}$	a set of
$\in$	element of
$\notin$	not an element of

$\|\ \|$	Magnitude of
$\mathcal{P}[\,]$	Phase of
$\mathcal{R}[\,]$	Real part of
$\mathcal{I}[\,]$	Imaginary part of
$\mathcal{Z}[\,]$	The zeros of

$f(x,y)$	the true image
$ff(x,y)$	autocorrelation of $f(x,y)$
$g(x,y)$	a convolution
$h(x,y)$	point spread function
$s(x,y)$	symmetric image
$S_f(x,y)$	the region of image space where $f(x,y)$ is non-zero
$B_f(x,y)$	the rectangular box, with sides parallel to the Cartesian coordinate system, just enclosing $S_f(x,y)$

CPE	Crude Phase Estimation
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
IFFT	Inverse Fast Fourier Transform
psf	point spread function
SAA	shift-and-add
ZAA	zero-and-add

Contents

Chapter 1

PRELIMINARIES

The purpose of this chapter is to introduce the terminology used and the mathematical techniques invoked in this thesis, whilst the practical applications of these techniques are discussed in the next chapter. The Fourier transform is fundamental to the algorithms presented in this thesis and is thus introduced in its dual roles as a mathematical model and computational tool. This chapter concludes with a description of the implementation of the Fourier transform on a digital computer.

It is well known that a unique relationship exists between an object and its expression in terms of its Fourier transform. This one-to-one invertible relationship between an object and its Fourier transform is introduced in §1.1. A computational analogy can be drawn between the use of the Fourier transform and the use of logarithms, both in the simplification of numerical processing and the modelling of physical phenomena. Just as logarithms are used to simplify multiplication by transforming it into addition, the Fourier transform reduces the process of convolution, discussed in §1.2, to multiplication. Logarithms can also be invoked when modelling the response of the human ear to sound intensity, which parallels the use of the Fourier transform in modelling the scalar wave equation (as shown in §1.3). The remainder of chapter 1 deals with the mathematical descriptions of the important physical situations in which the results of this thesis can be applied. These descriptions are, in general, unavoidably idealised but nevertheless have widespread applicability in many practical contexts. The notion of a compact image, §1.4, leads logically to sampled representations of the Fourier transform of an object, §1.5. §1.6 deals with the approximations needed to implement the Fourier transform on a digital computer.

In order to obtain wanted information from observable data it is necessary to formulate a mathematical model of any real world phenomenon. It is then required to represent this mathematical model in the form of a representation scheme (refer to Fig 1.1). Due to the widespread availability and convenience of digital computers this representation scheme is usually a finite set of numbers. The limitations of the mathematical modelling process have been addressed in the literature by a number of authors, notably Slepian (1983) and Requicha (1980), and it is inevitable that some of a model's faithfulness must be sacrificed in the interests of simplicity and computability. It is also essential that the overall process of converting a physical phenomenon into a representation scheme must be invertible, in order that calculations made using the representation scheme can be used to predict the behaviour of the physical process.

The usefulness of a model derives mainly from its ability to predict data that are unobservable from measurable quantities. It can also be used to condense large

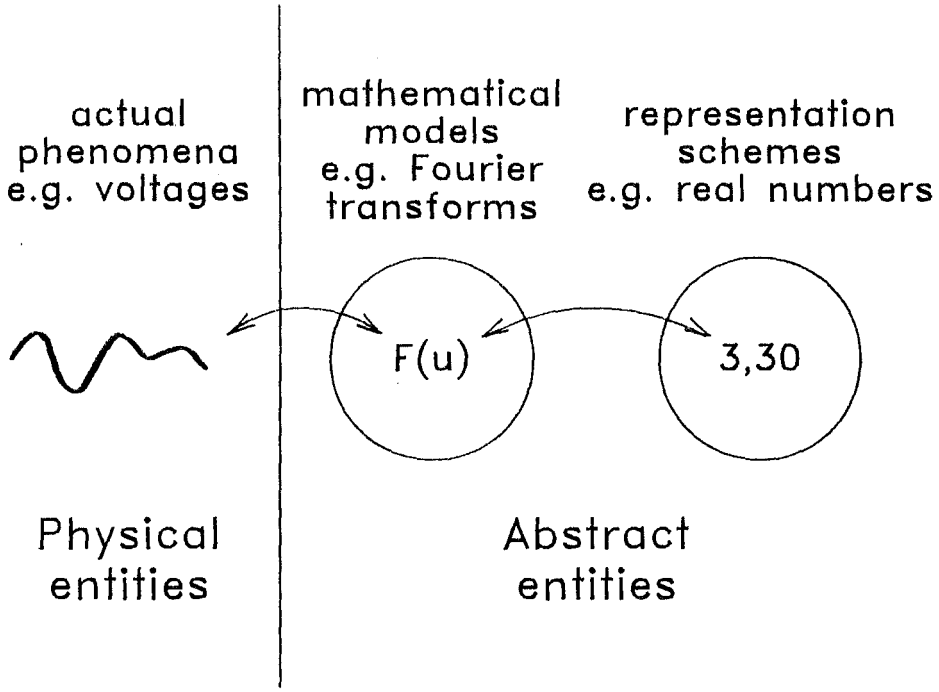


Figure 1.1: The relationship between models, representations and physical phenomena (after Requicha 1980)

quantities of data in the form of a few model parameters. Often a single mathematical model is utilised in a number of seemingly disparate areas. The Fourier transform (or its discrete analogue the Z-transform) provides an example of a widely applicable mathematical model, and much of the work in this thesis relies heavily on this mathematical formulation. An object's Fourier transform provides an alternative description related by a one-to-one invertible transformation between an object's description in conventional spatial coordinates and its description in terms of spatial frequencies.

Although complete knowledge of data related to an object's Fourier transform contains information equivalent to direct spatial measurement of the object, a display of the Fourier transform bears no visual resemblance to the original object. There are however a number of situations where the physical measurements which can be made correspond to the Fourier transform of the image (Table 1.1). Since these measurements relate to the visible portion of an object, the term visibility has come to be used for the Fourier transform of an object (Bates and McDonnell 1986). In this thesis both terms are used interchangeably.

An image can also be modelled with a sampled form of the Fourier transform. The sampled form of the Fourier transform is equivalent to the Z-transform which is discussed in chapter 3. It is, however, convenient to introduce the Z-transform notation along with that of the Fourier transform.

In this thesis an image, its Fourier spectrum (or alternatively visibility spectrum) and its Z-spectrum are said to exist in coordinate spaces known as image space, Fourier

Application	Object	Visibility	Image
X-ray and neutron crystallography	electron density of molecular structure	Diffraction pattern	real and positive
microscopy (acoustic, light and electron)	transmissivity or reflectivity of specimen (complex quantities in general)	Back focal plane field	complex
interferometry and spectroscopy, optical astronomical speckle imaging, radio astronomical aperture synthesis	spatially incoherent radiating source distributions	interferometric visibility, spatial frequency	real and positive
radio engineering, ultrasonics, acoustics	coherently radiating or induced source distributions	far field (Fraunhofer radiating pattern)	complex
communications, speech processing	signal	temporal spectrum	real

Table 1.1: The occurrence of the Fourier transform (after Fright 1984)

Space	Dimension		
	1	2	> 2
Image	$f(x)$	$f(x,y)$	$f(\vec{x})$
Fourier (visibility)	$F(u)$	$F(u,v)$	$F(\vec{u})$
Z	$\mathcal{F}(\zeta)$	$\mathcal{F}(\zeta,\gamma)$	$\mathcal{F}(\vec{\zeta})$

Table 1.2: Examples of the notation used for describing images and image transforms

space and Z-space respectively. Image space, Fourier space and Z-space are K-dimensional spaces spanned by the cartesian position vectors $\vec{x}$, $\vec{u}$ and $\vec{\zeta}$. Images are denoted by lower case Roman letters with the corresponding upper case letters being used to denote their Fourier transforms (as shown in Fig 1.2). Thus $F(\vec{u})$ is the Fourier transform of $f(\vec{x})$. A transform pair is denoted, for example, by $f(\vec{x}) \longleftrightarrow F(\vec{u})$, a notation which emphasises the invertible relationship between an image and its transform. In the special cases where K = 1 or 2, the notation listed in Table 1.2 is employed.

Although the Fourier transform and the Z-transform are denoted by the same upper case Roman letter, it is important to realise that whilst the former is usually a function of real coordinates $\vec{u}$, the latter is a function of complex coordinates $\vec{\zeta}$. It is often convenient, however, to analytically continue $\vec{u}$ into a complex vector in a manner discussed in chapter 3. Since the process of analytic continuation does not change the form of $F(\vec{u})$ when $\vec{u}$ is real, the above notation is also used for the analytically continued Fourier transform. It is important to realise that nearly all measurements correspond to when $\vec{u}$ is real.

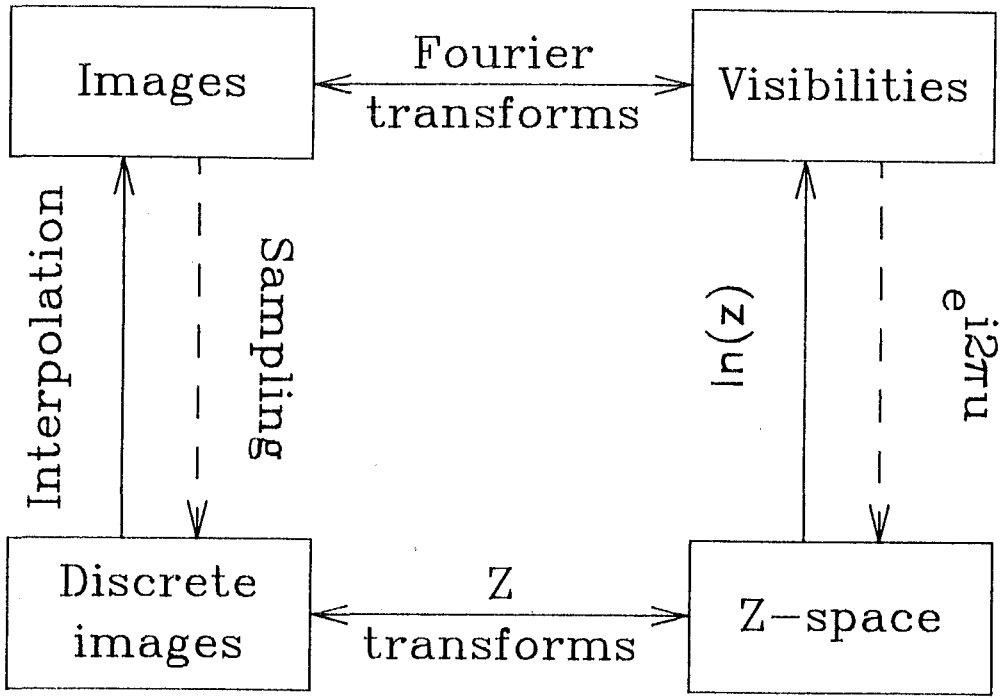


Figure 1.2: The interrelationship of image, Fourier and Z space.

In general, both the image and its transforms are complex, possessing both magnitudes and phases. A magnitude is denoted by $|\ |$, as in $|F(u)|$ and a phase by $\mathcal{P}[\ ]$, as in $\mathcal{P}[F(u)]$. Thus

$$F(\vec{u}) = |F(\vec{u})|e^{+i2\pi\mathcal{P}[F(\vec{u})]} \tag{1.1}$$

where i is the pure imaginary $= \sqrt{-1}$. Similarly the real and imaginary parts of, for example, $f(x)$ are denoted by $\mathcal{R}[f(x)]$ and $\mathcal{I}[f(x)]$ respectively, i.e.

$$f(\vec{x}) = \mathcal{R}[f(\vec{x})] + i\mathcal{I}[f(\vec{x})] \tag{1.2}$$

1.1 Fourier transforms

An image is related to its Fourier transform by

$$F(\vec{u}) = \int\!(K)\!\int f(\vec{x})e^{i2\pi\vec{u}\cdot\vec{x}}\, d\sigma(\vec{x}) \tag{1.3}$$

where $\int\!(K)\!\int$ denotes a K-dimensional integral, $d\sigma(x)$ is the K-dimensional volume element and

$$\vec{u}\cdot\vec{x} = \sum_{k=1}^{K} u_k x_k \tag{1.4}$$

As the images dealt with in this thesis are all of finite size and energy there is in general no difficulty in computing their Fourier transforms (Bates and McDonnell 1986, §6). The inverse Fourier transform is defined by

$$f(\vec{x}) = \int\!(K)\!\int F(\vec{u})e^{-i2\pi\vec{u}\cdot\vec{x}}\, d\sigma(\vec{u}) \tag{1.5}$$

and it is worth reiterating at this stage that (1.3) and (1.5) define a unique invertible relationship, provided the integrals are convergent (Bates and McDonnell 1986, §6). A

Fourier transform pair is denoted in this thesis by $f(\vec{x}) \longleftrightarrow F(\vec{u})$. This notation emphasises the close relationship between the forward and reverse Fourier transforms, in fact the choice of whether (1.3) is deemed the forward or reverse transform is purely a matter of convention (Bracewell 1978).

1.2 Modelling of scalar wave equations using the Fourier transform

The Fourier transform occurs naturally in many situations involving wave motion, where it can be used to relate the fields observed on parallel planes in a source free region. In particular, an exact Fourier transform exists between a source of finite size (or extent) and observations made at a plane at a distance much greater than the size of the source. The following discussion is by no means rigorous but summarises some of the situations where the Fourier transform is applicable. More complete derivations can be found in many works dealing with optics (cf. Born and Wolf 1970, Goodman 1968) or antenna theory (Silver 1965).

Consider a finite region of homogeneous space containing sources of linear wave motion, henceforth called the source distribution $\Upsilon_s(x, y, z, t)$. An arbitrary point P is chosen to lie within the source region. The space is spanned by three-dimensional Cartesian coordinates, with planes or lines for which z is constant being termed transverse. The source is observed on a transverse plane defined by $z = z_o$, henceforth referred to as the observation plane (Fig 1.3).

Wave motion is characterised by a scalar wave function $\Psi(x, y, z, t)$. When the source is assumed to be spatially coherent (all points of the source radiate in unison) and monochromatic (all points of the source radiate at the same frequency) the scalar wave function can be written in the form

$$\Psi = \Psi(x, y, z)e^{i\omega t} \tag{1.6}$$

where ω is the angular frequency of the radiation. Other useful parameters are the wavenumber k, the wavelength λ, and the wave propagation speed c which are related to ω by

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{c} \tag{1.7}$$

As is conventional the harmonic time dependence is suppressed (but is understood) throughout the following discussion of the behaviour of monochromatic wave functions.

In order to develop the relationship between the source distribution and the field on the observation plane it is convenient to define an equivalent source plane, positioned at $z = z_e$, such that $\Upsilon_s(x, y, z, t)$ lies entirely in the region of space given by $z < z_e$. The wave function on the the equivalent source plane is denoted by $\Psi_e(x_e, y_e)$ whereas the wave function in the observation plane is denoted by $\Psi_o(x_o, y_o)$, also known as the diffracted field.

It is possible by using Huygens' principle as expressed through Green's theorems (Silver 1965, §4.1) to define an equivalent two-dimensional source distribution $\Upsilon_e(x_e, y_e)$, in such a way that for $z > z_e$ the wave function emanating from the equivalent source is identical to the wave function resulting from the original source. As all practical imaging instruments are of finite size, only a finite region of the observation plane can be viewed.

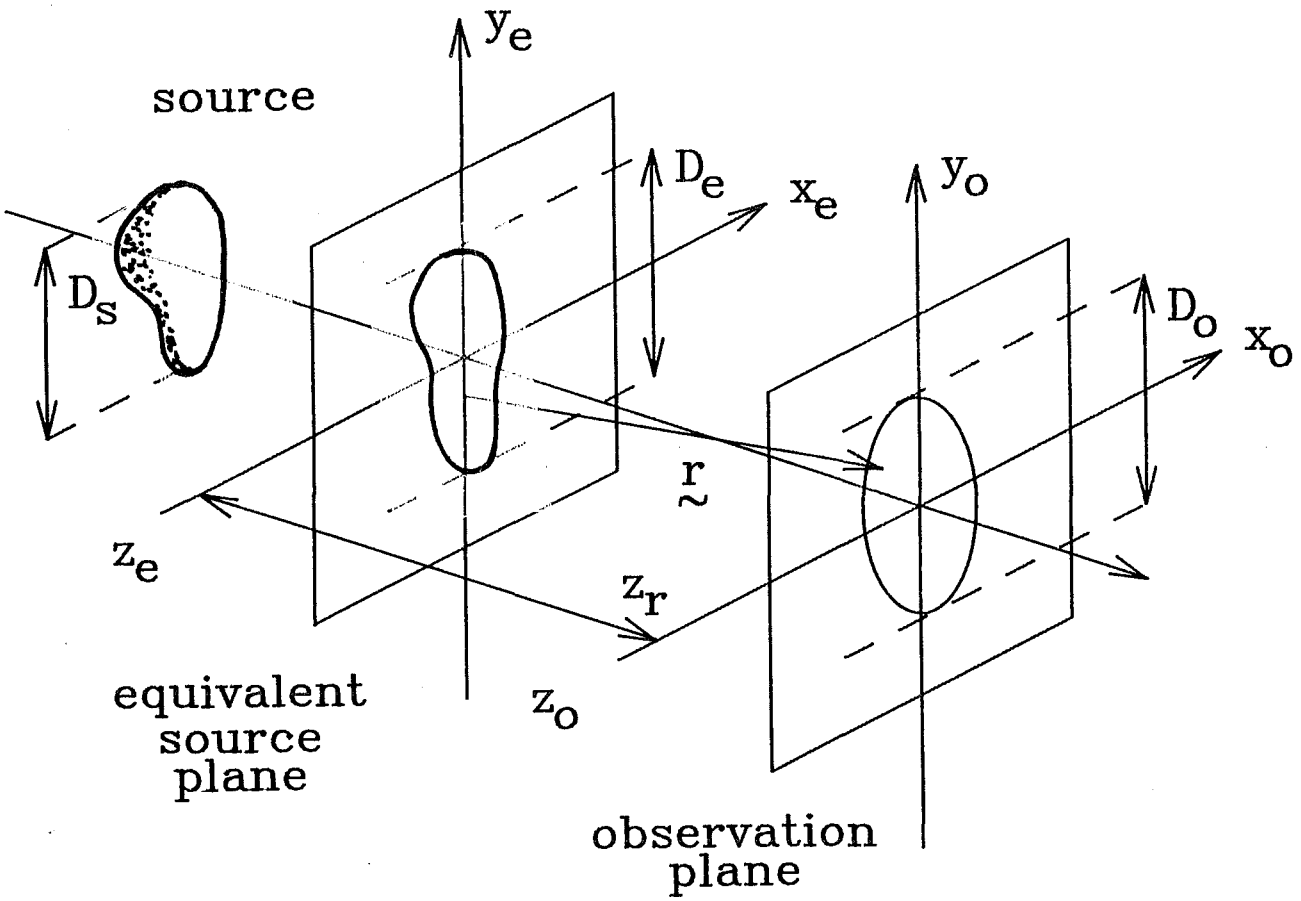


Figure 1.3: The source, equivalent source and observation planes

This finite planar region, or field of view, is determined by the size of the instrument's pupil, or aperture, and hence the observation plane is often called the aperture plane. The largest transverse dimension of the aperture is denoted by D_o. The size of the aperture limits the largest transverse dimension of the equivalent two-dimensional source, since rays from the arbitrary source point P intersecting the field of view can only intersect the equivalent source plane over a finite area, which has a largest transverse dimension denoted by D_e (Silver 1965, §5.11). Although D_e must be larger than the largest transverse dimension of the actual source D_s , if the distance $(z_o - z_e)$ is very large, then

$$D_e \approx D_s \tag{1.8}$$

The use of equivalent sources in astronomy provides a very effective and very ancient model (Bates 1982a, §2.2). The modelling of stars as two-dimensional sources on a celestial sphere of vast radius dates to the origins of astronomy. Having defined the equivalent source and observation planes it is possible to express the diffracted field, or visibility, in terms of the equivalent source field by using the Rayleigh-Sommerfield formula (Goodman 1968, p 44):

$$\Psi_o(x_o, y_o) = \frac{1}{i\lambda} \iint \Psi_e(x_e, y_e) \cdot e^{ik|\vec{r}|} \cdot \frac{cos(\vec{z}, \vec{r})}{|\vec{r}|} dx_e dy_e \tag{1.9}$$

where $\vec{r}$ is the vector between points in the equivalent source and points within the aperture, and $\cos(\vec{z}, \vec{r})$ is the cosine of the angle between $\vec{r}$ and the z axis (Fig 1.3). z_r is the z coordinate of r and is thus equal to $(z_o - z_e)$.

By assuming that

$$|\vec{r}| \gg D_s \text{ or } D_e \tag{1.10}$$

it is reasonable to approximate $\cos(\vec{z},\vec{r})$ by unity and $1/|\vec{r}|$ by $1/z_r$. In many cases, however, k can be very large and it is necessary to employ a more accurate estimate of $|\vec{r}|$ in the $e^{(ik|\vec{r}|)}$ term. The exact expression for $|\vec{r}|$ is

$$|\vec{r}| = \sqrt{z_r^2 + (x_e - x_o)^2 + (y_e - y_o)^2} \tag{1.11}$$

which can be expanded using the binomial theorem. Discarding terms of higher than first order yields

$$|\vec{r}| \approx z_r + \frac{(x_e - x_o)^2 + (y_e - y_o)^2}{2} \tag{1.12}$$

which when substituted in (1.9), yields the Fresnel or near-field approximation (Goodman, §4.1):

$$\Psi_o(x_o, y_o) = \frac{e^{ikz_r}}{i\lambda z_r} \cdot e^{\frac{ik(x_o^2+y_o^2)}{2z_r}} \iint \Psi_e(x_e, y_e) \cdot e^{\frac{ik(x_e^2+y_e^2)}{2z_r}} \cdot e^{\frac{i2\pi(x_e x_o + y_e y_o)}{\lambda z_r}} dx_e dy_e \tag{1.13}$$

When $|\vec{r}|$ is large enough that the quadratic terms of the Fresnel transform can be discarded resulting in the Fraunhofer or far-field approximation (Goodman §4.1):

$$\Psi_o(x_o, y_o) = \frac{e^{ikz_r}}{i\lambda z_r} \cdot e^{\frac{ik(x_o^2+y_o^2)}{2z_r}} \iint \Psi_e(x_e, y_e) \cdot e^{\frac{i2\pi(x_e x_o + y_e y_o)}{\lambda z_r}} dx_e dy_e \tag{1.14}$$

The Fraunhofer approximation is usually applicable (Silver 1965, §6.9) in situations where

$$|\vec{r}| > \frac{2D_e^2}{\lambda} \tag{1.15}$$

By comparing the expression for the Fraunhofer approximation with the definition (1.3) of the Fourier transform, it can be seen that apart from a complex scale factor a Fourier transform relationship exists between the fields in the source and observation planes. The Fresnel transform, by contrast, requires the removal of the quadratic phase variation before a Fourier transform relationship can be applied.

The above derivation has dealt with monochromatic spatially incoherent sources. Also of interest are cases where the source is spatially incoherent, i.e. radiations from different points of the source are statistically independent. Bodies emitting thermal radiation, such as stars or incandescent lamps are examples of spatially incoherent sources (Bates 1982a). This is because groups of atoms or molecules separated by discernible distances do indeed radiate independently.

When points of the source are statistically independent the total response is obtained by summing the effects of individual points on the basis of their intensity. The visibility which is then observed is thus the Fourier transform of the intensity distribution of the source provided the wave motion is of a finite bandwidth $\Delta\omega$, which must be small compared to the mean frequency ω_m i.e.

$$\frac{\Delta\omega}{\omega_m} \ll 1 \tag{1.16}$$

(Bates and McDonnell 1986, §3 and §8). Wave motion in this case is known as quasi-monochromatic and the wave function now represents the intensity of wave motion in the band of width $\Delta\omega$ and centred on ω_m. It is important to note that the equivalent source is also an intensity distribution and as such is known to be positive (i.e. real and non-negative).

1.3 Convolutions

In the previous section the Fourier transform is shown to provide a model of scalar wave diffraction. In this section the versatility of the Fourier transform is emphasised, by showing the extent to which it simplifies the modelling of the process of convolution. The formation of an image of an object is often limited by an intervening distortion (Bates and McDonnell, §3). This distortion can be essentially random in nature, for example the twinkling of stars caused by atmospheric turbulence, or of a fixed form such as an out of focus camera. In the ensuing discussion the term system is loosely used to describe a physical process where all points in the undistorted image are modified in some way. In the following discussion, this distortion is denoted mathematically as $S[\ ]$, which in the important case of linear, point spread invariant systems is known as a convolution.

In modelling the process of convolution it is convenient to consider the most elementary possible object, the point source. This is an object which although visible is too small for any detail to be resolved. Stars are often modelled by point sources. This is because although physically very large they are, except for certain red giants in our galaxy, so distant that they cannot be resolved even in the largest of existing telescopes.

The point source is appropriately modelled by the Dirac delta function $\delta(\vec{x})$ (Bates and McDonnell 1986, §6) which is a function with non-zero value only for a particular value of $\vec{x}$. $\delta(\vec{x})$ is defined by

$$\begin{aligned} \int_{x\in\Lambda} \delta(\vec{x})\, d\sigma(\vec{x}) &= 1 \\ \int_{x\notin\Lambda} \delta(\vec{x})\, d\sigma(\vec{x}) &= 0 \end{aligned} \tag{1.17}$$

where Λ is an infinitesimally small region of $\vec{x}$-space containing $\vec{x} = 0$ The delta function has a number of useful mathematical properties of which the sifting property is perhaps the most useful

$$\int f(\vec{x})\delta(\vec{x} - \vec{x}')\, d\sigma(\vec{x}) = f(\vec{x}') \tag{1.18}$$

Any image can thus be considered to be a weighted sum of delta functions, e.g.

$$f(\vec{x}) = \int f(\vec{x}')\delta(\vec{x} - \vec{x}')\, d\sigma(\vec{x}) \tag{1.19}$$

Having defined an elemental function from which any image can be constituted it is useful to consider the response of a system to a single delta function

$$h(\vec{x}, \vec{x}') = S[\delta(\vec{x} - \vec{x}')] \tag{1.20}$$

This response is known as the point spread function (psf) of a system. If the system is linear it is possible to apply superposition to obtain the system output by integrating the response to the individual input delta functions. If $g(\vec{x})$ is the output of the linear system S[ ] to the input $f(\vec{x})$ then

$$\begin{aligned} g(\vec{x}) &= S[f(\vec{x})] \\ &= S[\int f(\vec{x}')\delta(\vec{x} - \vec{x}')\, d\sigma(\vec{x}')] \\ &= \int S[f(\vec{x}')\delta(\vec{x} - \vec{x}')\, d\sigma(\vec{x}')] \\ &= \int f(\vec{x}')h(\vec{x}, \vec{x}')\, d\sigma(\vec{x}') \end{aligned} \tag{1.21}$$

An important class of linear systems encompasses those for which $\vec{x}$ and $\vec{x}'$ are of the same dimension and the point spread function or psf $h(\vec{x}, \vec{x}')$ is only a function of the difference between $\vec{x}$ and $\vec{x}'$ i.e.

$$h(\vec{x}, \vec{x}') = h(\vec{x} - \vec{x}') \tag{1.22}$$

These systems are termed point spread invariant (psi) or isoplanatic, as the shape of the point spread function in x-space is not a function of $\vec{x}$. The well known convolutional integral (Bates and McDonnell 1986, §7) is obtained by substituting (1.22) into (1.21).

$$g(\vec{x}) = \int f(\vec{x}')h(\vec{x} - \vec{x}')\, d\sigma(\vec{x}') \tag{1.23}$$

or

$$g(\vec{x}) = f(\vec{x}) \odot h(\vec{x}) \tag{1.24}$$

where $\odot$ is the symbol that represents convolution. The images $f(\vec{x})$ and $h(\vec{x})$ are termed components of the convolution $g(\vec{x})$. Because (1.23) is inherently more mathematically tractable than (1.21) it is often used as an approximation when the psf varies slowly with $\vec{x}$ whereupon (1.23) is applied over small regions known as isoplanatic patches.

The applicability of the Fourier transform to psi systems arises because these systems have the complex exponentials $e^{i(\vec{u}_0 \cdot \vec{x})}$ as eigenfunctions (Slepian 1985). Thus if the input $f(\vec{x})$ is the complex exponential $e^{i\vec{u}_0 \cdot \vec{x}}$ then the output $g(\vec{x})$ is given by

$$g(\vec{x}) = S[e^{i\vec{u_o} \cdot \vec{x}}] = H(\vec{u_o}) \cdot e^{i\vec{u_o} \cdot \vec{x}} \tag{1.25}$$

Because the output and input only differ by a complex scale factor $H(\vec{u}_0)$, this scale factor must be an eigenvalue of the linear psi system defined by $S[\,]$. The complex exponential $e^{i(\vec{u}_0 \cdot \vec{x})}$ is the eigenvector corresponding to $H(\vec{u}_0)$. The Fourier transform can thus be used to obtain the eigenvalues of $S[\,]$ from $h(\vec{x})$ since the eigenvalue at any point $\vec{u_0}$ is given by $H(\vec{u_0})$.

When the Fourier transform is applied to (1.23) one obtains the convolution theorem (Bates and McDonnell 1986, §7)

$$G(\vec{u}) = F(\vec{u})H(\vec{u}) \tag{1.26}$$

This simple result is of deep significance providing a powerful tool for analysing the process of convolution. The Fourier transform converts the problem of deconvolution in image-space to one of factorisation in Fourier-space. Also of note in (1.26) is the symmetry between the psf and the input image. It is worth noting that it is not possible, from (1.26) alone, to distinguish whether $F(\vec{u})$ in fact corresponds to the input image or the psf without extra information (i.e. $g(\vec{x})$ could equally well result from inputting $h(\vec{x})$ to a linear psi system with a psf equal to $f(\vec{x})$).

An important example of a convolution is the autocorrelation of an image $f(\vec{x})$, which is here defined by

$$ff(\vec{x}) = f(\vec{x}) \odot f^*(-\vec{x}) \tag{1.27}$$

or from the autocorrelation theorem (Bates and McDonnell 1986, §7)

$$ff(\vec{x}) \longleftrightarrow |F(\vec{u})|^2 = F(\vec{u})F^*(\vec{u^*}) \tag{1.28}$$

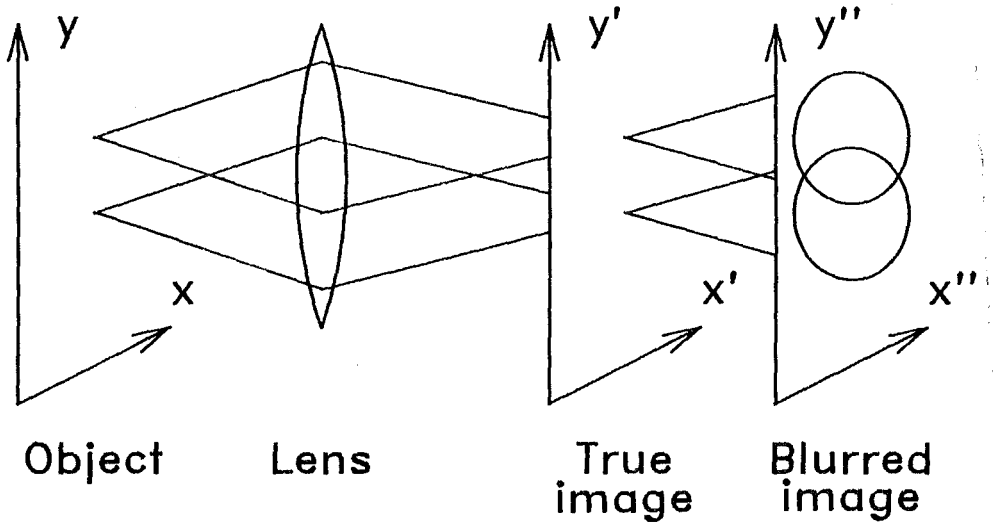


Figure 1.4: Schematic illustration of an out of focus camera

The autocorrelation can thus be calculated, using (1.28), solely from the visibility magnitude. It is therefore relevant in situations where the visibility phase cannot be measured (§2.4). A familiar example of deconvolution is provided by an out of focus camera, as illustrated in Fig 1.4. The result of the blurring is to map each point of the ideal (i.e. undistorted) image onto a disc in the output image. This has the effect of averaging or smearing the image as shown in Fig 1.5.

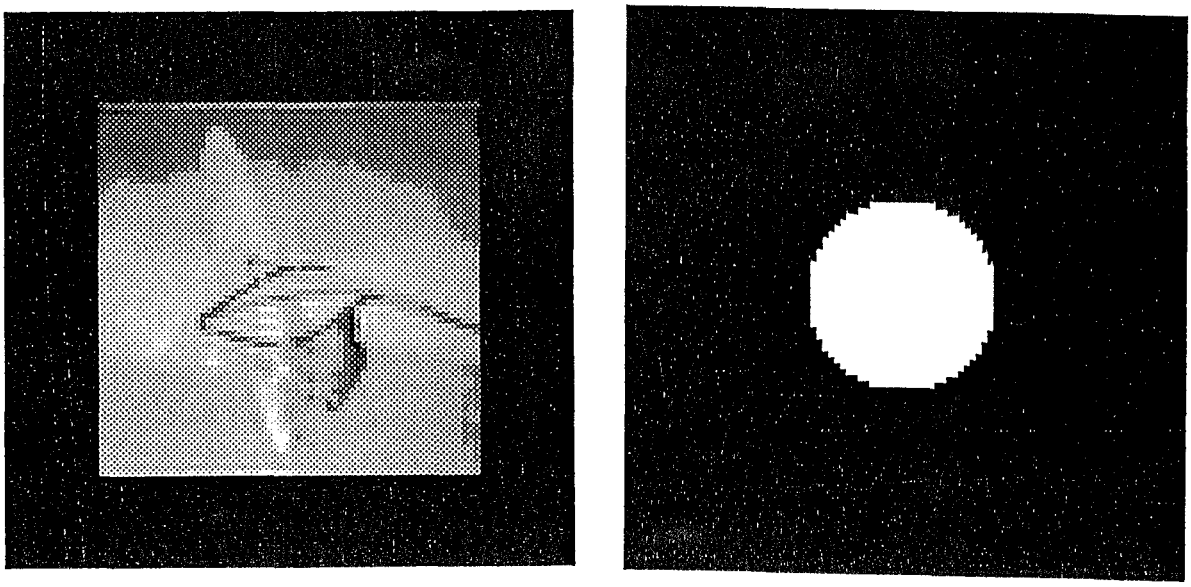
1.4 Compact Images

All real world images are effectively of finite size and amplitude. Slepian (1983) notes that the concept of infinite size or infinite amplitude is more often than not a product of mathematical abstraction rather than physical reality and is in practice physically unverifiable. Although the following discussion is in terms of compactness in image-space, the concept can be applied to an object's Fourier transform simply by replacing $f(\vec{x})$ with $F(\vec{u})$.

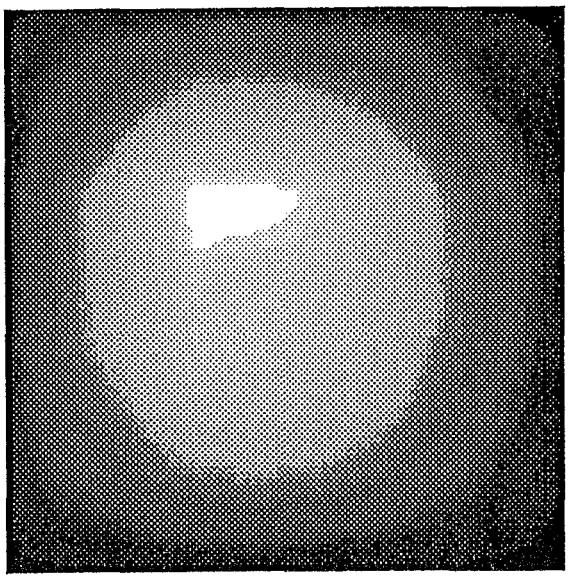
As a consequence of being of finite energy, most of the energy of an image $f(\vec{x})$ can be contained in a finite region of space image space known as the support. The support of $f(\vec{x})$ is denoted by $S_f(\vec{x})$. An image compact in image space satisfies (by definition)

$$\frac{\int_{S_f(\vec{x})} |f(\vec{x})|^2 d\sigma(\vec{x})}{\int_{\vec{x}} |f(\vec{x})|^2 d\sigma(\vec{x})} \geq 1 - \xi^2 \tag{1.29}$$

where ξ^2 determines the ratio of the image energy outside the support to the total image energy. An image is thus not necessarily identically zero outside its support although, as is discussed in §1.6, it may prove convenient to assume so. In practice ξ^2 is determined by when the magnitude of an image becomes comparable with the background noise or uncertainty level in the measurements.



(a) (b)



(c)

Figure 1.5: Demonstration of the blurring caused by an out of focus camera. Note that the convolution is larger than the input image.
(a) the input image, $f(x, y)$
(b) the psf, $h(x, y)$
(b) the blurred output image, $g(x, y) = f(x, y) \odot h(x, y)$

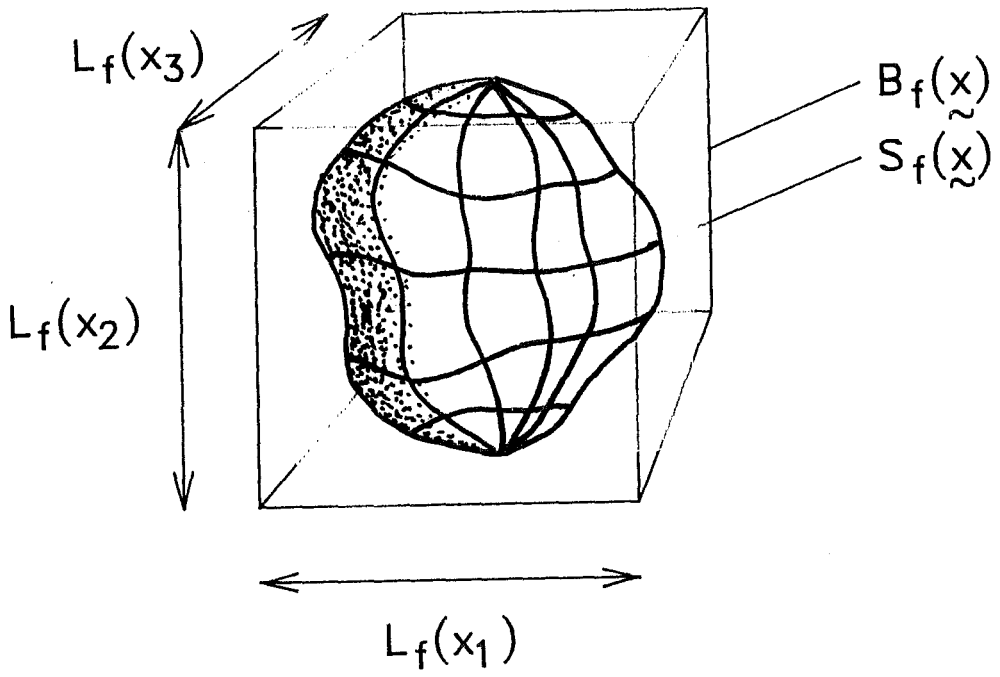


Figure 1.6: The support and image box for an image $f(\vec{x})$

A compact image is also of finite amplitude i.e.

$$|f(\vec{x})| < \infty \quad \forall \vec{x} \tag{1.30}$$

It is often convenient to define an image-box $B_f(\vec{x})$ which is the region just enclosing the support of $f(\vec{x})$ with sides parallel to arbitrarily chosen Cartesian coordinates in image space (cf Table 1.2). The image-box of an image is thus always larger than or equal to the support of the image (Bates 1982b).

$$S_f(\vec{x}) \subset B_f(\vec{x}) \tag{1.31}$$

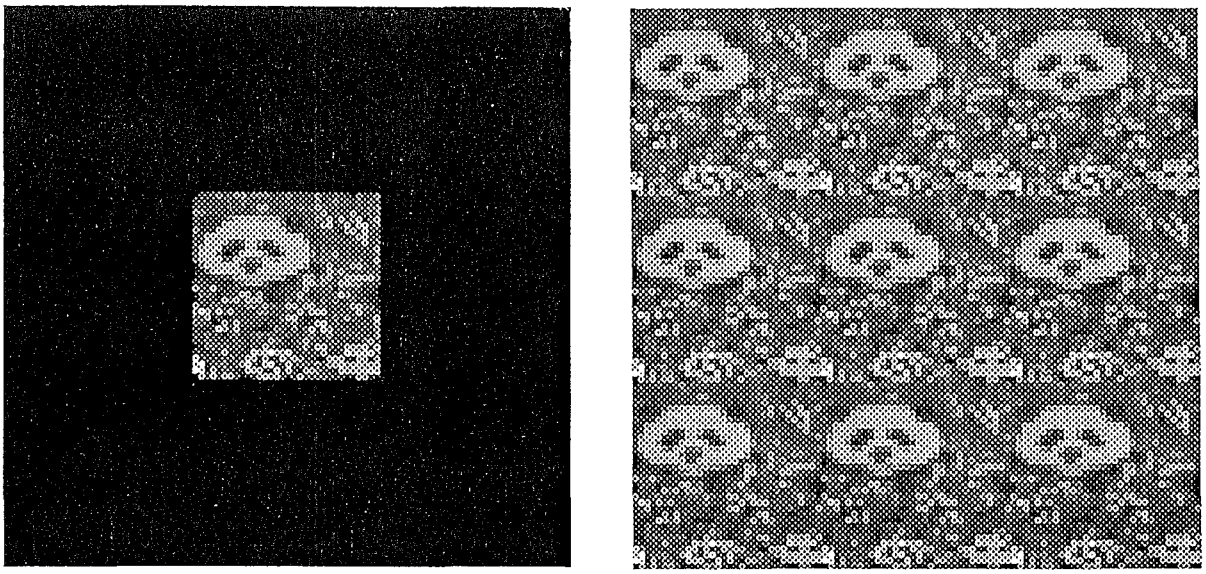
The image-box extent, or linear dimension of the image-box in terms of any specific Cartesian coordinate, is denoted by $L_f(x_k)$. Fig 1.6 shows an example of an image support and image-box.

A compact image can be represented to an arbitrary accuracy within its support by a truncated Fourier trigonometric series, referred to hereafter as a finite Fourier series. The finite Fourier series representation of $f(\vec{x})$ is here written as $p(\vec{x})$, which is defined by

$$p(\vec{x}) = \sum_{k=1}^{K} \sum_{m_k=-M_k}^{M_k} F(m_1,,,m_k) e^{i2\pi\left(\frac{m_1 x_1}{L_f(x_1)}+\cdots+\frac{m_K x_K}{L_f(x_K)}\right)} \tag{1.32}$$

where $F(m_1,,,m_K)$ are complex constants given by

$$F(m_1,,,m_k) = \frac{\int(K)\int f(\vec{x}) e^{-i2\pi\left(\frac{m_1 x_1}{L_f(x_1)}+\cdots+\frac{m_K x_K}{L_f(x_K)}\right)} d\nabla(\vec{x})}{\int(K)\int_{B_f(\vec{x})} d\sigma(\vec{x})} \tag{1.33}$$



(a) (b)

Figure 1.7: Illustration of the difference between $f(x,y)$ and $p(x,y)$
(a) $f(x,y)$ (b) $p(x,y)$

and the accuracy of the representation within the support can be calculated using,

$$\epsilon^2 = \frac{\int_{S_f(\vec{x})} |f(\vec{x}) - p(\vec{x})|^2}{\int_{\vec{x}} d\sigma(\vec{x}) \, |f(\vec{x})|^2} \tag{1.34}$$

It is apparent from (1.32) that $p(\vec{x})$ has a period of $L_f(x_k)$ in the x_k direction. It is therefore only an accurate model of $f(\vec{x})$ within the support of $f(\vec{x})$. Outside $B_f(\vec{x})$, where $f(\vec{x})$ is actually small, $p(\vec{x})$ replicates $f(\vec{x})$ as is illustrated in Fig 1.7.

In the absence of noise, the accuracy of the representation of a compact image by a Fourier series is mainly a function of the number of harmonics used, as illustrated by Fig 1.8. It should be noted that, although the error defined in (1.34) declines to zero with increasing numbers of harmonics, the absolute error

$$\epsilon_a = \max_{S_f(\vec{x})} |f(\vec{x}) - p(\vec{x})| \tag{1.35}$$

does not, a result of the well known Gibb's phenomenon (Kreysig 1979, p 504). The use of trigonometric polynomials to model images which are corrupted by noise is discussed in chapter 3. In such cases only a certain number of harmonics can be measured and as a result ϵ, as defined in (1.34), can actually increase when the number of harmonics used exceeds a limit determined by the level of the noise (Napier 1971).

1.5 Sampling

In §1.1 the concept of a representation scheme is introduced, providing a method of representing a mathematical model numerically. Sampling, the description of a continuous

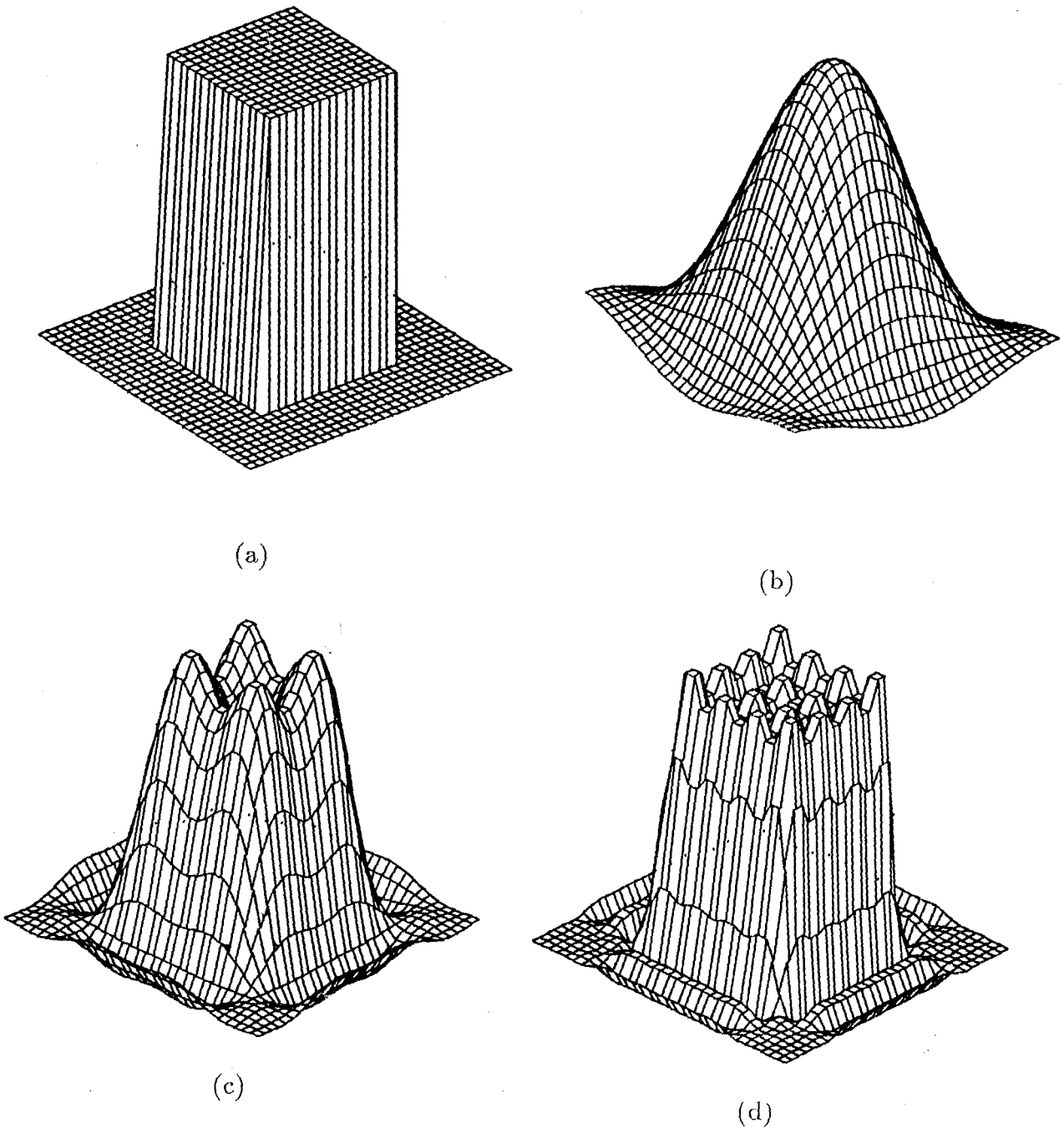


Figure 1.8: Demonstration of how an image is made up from harmonics. The number of harmonics used is determined by M_k in (1.32).
(a) $f(x,y)$ (b) $p(x,y)$ when $M_k = 2$ (c) $p(x,y)$ when $M_k = 4$ (d) $p(x,y)$ when $M_k = 8$

object by its amplitude at discrete points, is in general an essential step in reducing a mathematical model to a computable representation scheme. The process of sampling is inherent in all digital processing.

Examination of (1.32) shows that $p(\vec{x})$ is completely specified by the set of complex constants $F(m_1...m_K)$ in (1.33). Comparing (1.32) and (1.5) shows that these constants are in fact samples of the $F(\vec{u})$ taken on a K-dimensional grid. These samples are spaced in the k^{th} dimension by ε_k, where

$$\varepsilon_k = \frac{1}{L_f(x_k)} \tag{1.36}$$

which is known as the Whittaker, Shannon or Nyquist sampling. An example of a practical periodic object is a crystal structure for which the Fourier transform exists only at these samples or structure factors (Ramachandran and Srinivasan 1970). In the case of an aperiodic compact image the Fourier transform is continuous and the sampling is thus arbitrary. There is an important relationship between extent in image space and sampling in Fourier space. If the sampling rate is greater than the Nyquist rate then $p(\vec{x})$ consists of repeated versions of $f(\vec{x})$ surrounded by empty space, a process often referred to as packing with zeros (see Fig 1.9b). The zero-packed $f(\vec{x})$ denoted hereafter as $f'(\vec{x})$ has an extent increased in proportion to the amount of oversampling, i.e.

$$\frac{\varepsilon'_k}{\varepsilon_k} = \left\lceil \frac{L'_f(x_k)}{L_f(x_k)} \right\rceil^{-1} \tag{1.37}$$

If Fourier-space is sampled at a rate lower than the Nyquist rate then $p(\vec{x})$ again consists of repeated versions of $f(\vec{x})$, but they are now overlapped due to a process known as aliasing Fig 1.9c. The period of $p(\vec{x})$ is now less than the extent of $f(\vec{x})$ and information is lost as it is no longer possible to recover $f(\vec{x})$ from $p(\vec{x})$ uniquely by isolating a single period of $p(\vec{x})$.

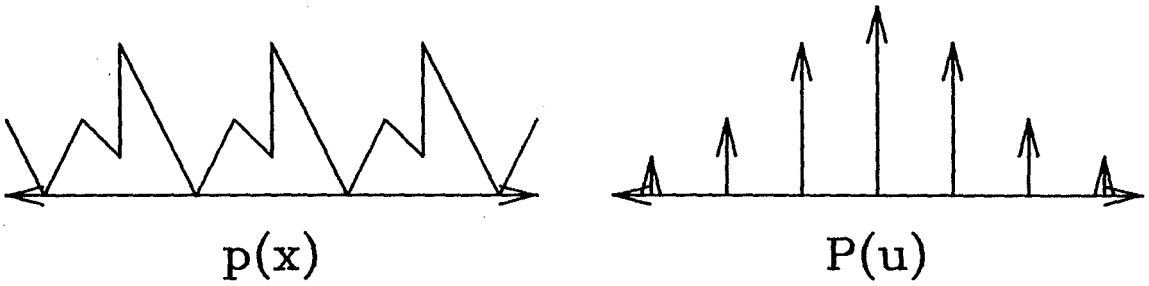
The convolution of compact images can be seen from (1.23) to produce a convolution of greater extent than either of the two convolved images. More specifically the extent of the convolution of two images $f(\vec{x})$ and $g(\vec{x})$ is no greater than the sum of the extents of $f(\vec{x})$ and $g(\vec{x})$ (Bates and McDonnell 1978, §8). Thus

$$L_{f \odot g}(x_k) \leq L_f(x_k) + L_g(x_k) \tag{1.38}$$

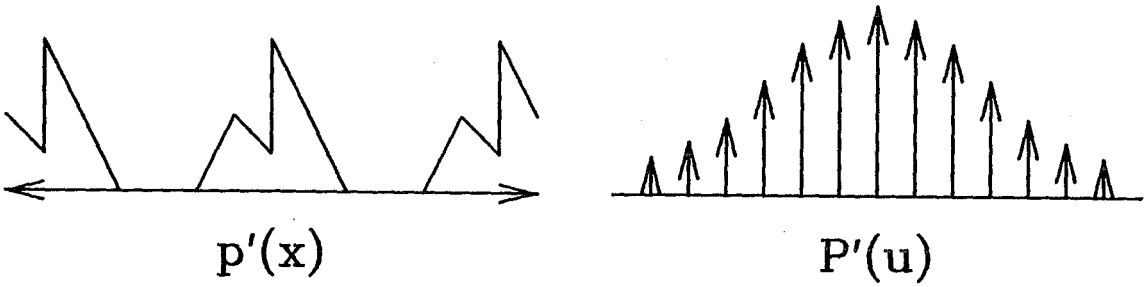
where equality holds when $f(x,y)$ and $g(x,y)$ are positive images. Fig 1.5 shows the convolution of two compact components. The increased size of a convolution, when compared to the size of its components, is readily apparent. Because of the connection between sampling in image space and the extent in Fourier space (1.36), a higher sampling rate is required in Fourier space in order to adequately represent the convolution's visibility. If when calculating the convolution from (1.26) the visibilities of the images are not sampled at the Nyquist rate of the convolution, then the calculated convolution suffers from aliasing. In other words, when using (1.26) to calculate a convolution, the components of the convolution must be zero packed to the size of the resultant convolution, in order to increase their sampling rate in Fourier space.

1.6 Compactness in Fourier- and image-space

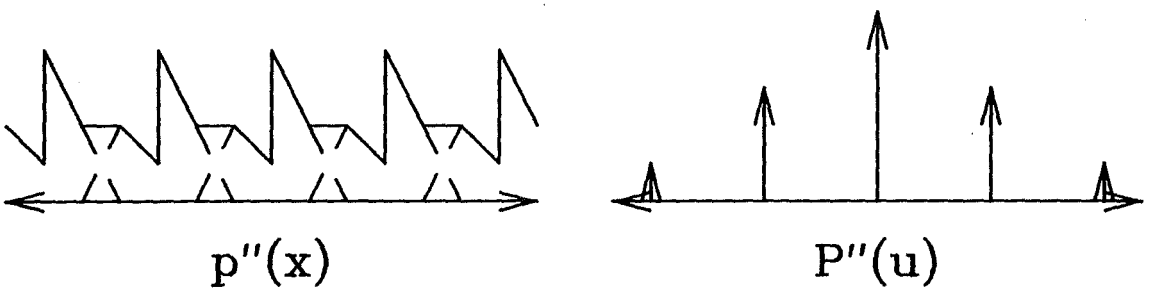
Although, in the description of compact images in §1.4, it was stated that compactness is only relative to a given level determined by the accuracy of measurement, it is useful



(a)



(b)



(c)

Figure 1.9: Relationship between sampling rate in Fourier space and periodicity in image space
(a) Nyquist sampling
(b) Zero packing in image space $\longleftrightarrow$ oversampling in Fourier space
(c) Aliasing in image space $\longleftrightarrow$ undersampling in Fourier space

to introduce the concept of an exactly compact image. An exactly compact image is a mathematical abstraction defined by letting $\xi = 0$ in (1.29). The Fourier transform of an exactly compact image must be of infinite extent (Bates and McDonnell 1986, Bracewell 1978). In reality, it is as difficult to confirm whether an image is of infinite extent in Fourier-space, as it is to verify whether the image is precisely zero outside a given region of image-space.

Nearly all real world objects are of finite energy. As a result of Parseval's theorem (Bracewell 1978, called the energy conservation theorem by Bates and McDonnell 1986), this energy can be calculated in either image-space or Fourier-space:

$$\int_{\vec{x}} |f(\vec{x})|^2 d\sigma(\vec{x}) = \int_{\vec{u}} |F(\vec{u})|^2 d\sigma(\vec{u}) \tag{1.39}$$

A consequence of (1.39) is that all real world objects are effectively compact in both image-space and Fourier-space. It is, however, often useful to make the assumption that an object is exactly compact in, for example, image-space, to facilitate mathematical analysis of physical situations. As has been noted by Slepian (1983) this is quite acceptable as long as the results derived from the ensuing model are not sensitive to the assumption of exact compactness.

The following is a brief summary of some classic papers (Slepian 1983, Landau and Pollack 1962 and 1961, Slepian and Pollack 1961, Papoulis 1984). It has been necessary, however, to alter the notation of these references in order to maintain consistency with the form of the Fourier transform defined by (1.3). The analysis of a one-dimensional function compact in Fourier space begins by assuming the visibility is exactly compact within $[-W, W]$ i.e.

$$|F(\vec{u})| \begin{cases} \geq 0 & for \quad |u| \leq W \\ = 0 & for \quad |u| > W \end{cases} \tag{1.40}$$

There is then a tradeoff to be made in image space between how compact the image is in the interval $[-X/2, X/2]$, i.e.

$$\frac{\int_{\frac{-X}{2}}^{\frac{+X}{2}} |f(\vec{x})|^2 dx}{\int_{-\infty}^{+\infty} |f(\vec{x})|^2 dx} = 1 - \xi^2 \tag{1.41}$$

and how accurately it is approximated within the support

$$\inf_{\{a_i\}} \int_{\frac{-X}{2}}^{\frac{+X}{2}} |f(\vec{x}) - \sum_{0}^{N} a_n g_n(\vec{x})|^2 dx < K\xi^2 \tag{1.42}$$

where $g_n(x)$ are basis functions chosen so that the $G_n(u)$ are also exactly compact in the interval $[-W, W]$. Landau and Pollack (1962) find that if the $g_n(x)$ are chosen to be the prolate spheroidal wavefunctions (pswfs) $\Phi_n(x)$ then (1.42) is satisfied for $N = 2WX$ and $K = 12$. These functions are defined by the integral equation (Slepian 1983)

$$\lambda_n \Phi_n(x) = \int_{\frac{-X}{2}}^{\frac{+X}{2}} \Phi_n(s) \frac{sin(2\pi W(x-s))}{\pi(x-s)} ds \tag{1.43}$$

and form a set of functions orthogonal in the interval [-X/2,X/2]. The concentration of $\Phi_n(x)$ within [-X/2,X/2] is given by

$$\lambda_n = 1 - \xi_n^2 \tag{1.44}$$

and this falls rapidly for $n > 2WX$. Thus the $\Phi_n(x)$ are only effectively compact for n in the range $[0, 2WX]$. Using more than the first $2WX + 1$ functions improves the accuracy to which the function is approximated within $[-X/2, X/2]$, but also causes the function to assume large values outside the support. As the $\Phi_n(x)$ are determined solely by the product WX (Slepian 1983), only the $(2WX + 1)$ a_n can be varied to give different $f(x)$. The dimension of the function space of one-dimensional functions limited to $[-X/2, X/2]$ in image space and $[-W, W]$ in Fourier space is thus essentially $2WX + 1$ dimensional.

Other popular basis functions are the sampling functions

$$samp(k) = \frac{sin(\pi(2WX - k)}{\pi(2WX - k)} \tag{1.45}$$

whereupon the approximation error is given by

$$\int_{-\frac{X}{2}}^{+\frac{X}{2}} |f(x) - \sum_{0}^{N} f\left[\frac{k}{2W}\right] samp(k)|^2 dx \leq K\xi^2 \tag{1.46}$$

Although the optimal coefficients of the sampling functions are easily calculated, since the $(f(k/2W)$ are simply samples of the image, they cannot satisfy (1.46) independently of ξ^2 for any fixed value K. In order to adequately represent some $f(x)$ using sampling functions it is necessary to sample at a rate higher than the Nyquist rate. As sampling in image space determines the extent in Fourier space (Bates and McDonnell 1986), oversampling the image no longer constrains the visibility to be compact in the interval $[-W, W]$. Representing a function in terms of a number (of basis functions) which is more than the dimension of the space they span is a well known source of ill-conditioning (Landau and Pollack 1962). This problem is discussed further in chapter 7.

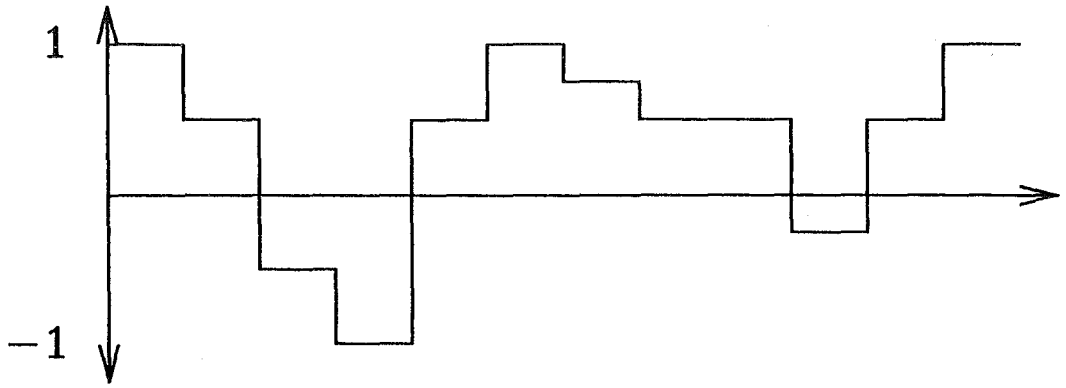
1.7 The discrete Fourier transform

The process of sampling implicitly bandlimits the Fourier transform because frequencies higher than the Nyquist frequency are aliased. When it is necessary to sample the Fourier transform as well, the image is also forced to be compact. The previous section has discussed how the space of images which are simultaneously compact in both Fourier and image space is effectively of dimension $2WX + 1$, although more than $2WX$ samples are required to adequately represent an image. In light of the resultant ill-conditioning it may appear inappropriate that the sampling functions are still preferred to pswfs to represent an image. In practice, however, the continuous pswfs are numerically inconvenient, whilst the sampled image and its visibility can be efficiently related using the DFT (the discrete pswfs are another option and are discussed in chapter 5). The one-dimensional DFT is given here for simplicity although the K-dimensional extension is straightforward.

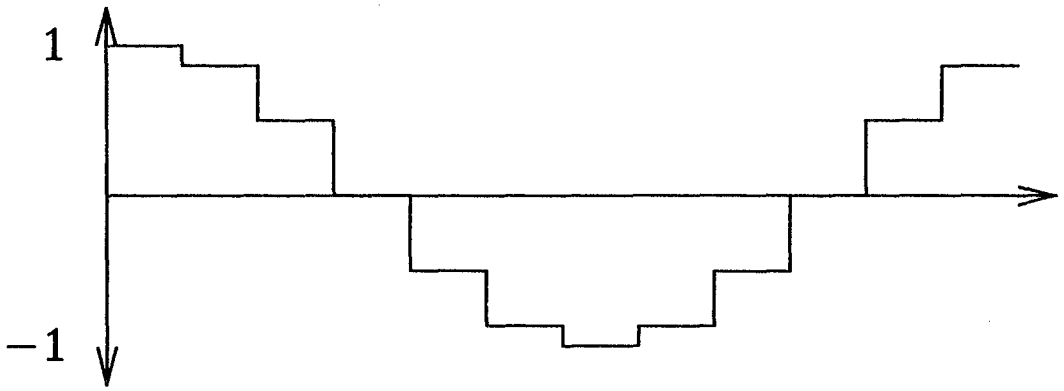
$$F_m = \sum_{n=0}^{N-1} f_n e^{\left(\frac{i2\pi mn}{N}\right)} \tag{1.47}$$

The inverse DFT is defined by

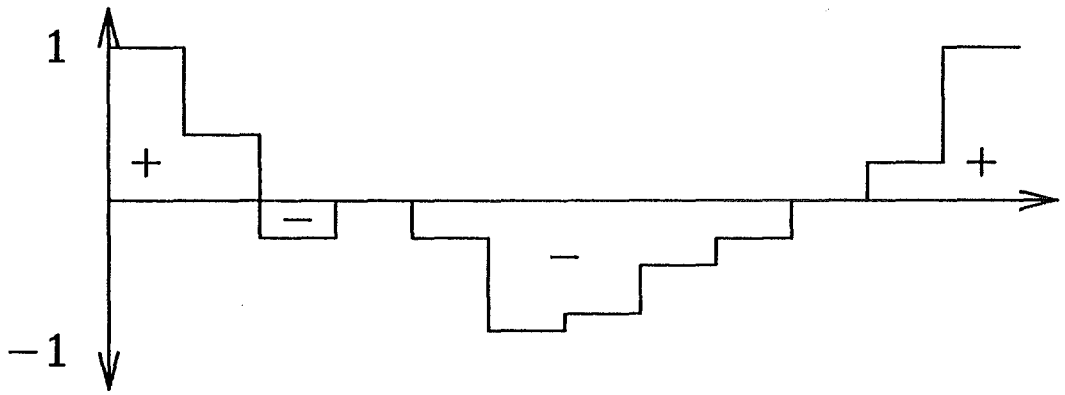
$$f_n = \frac{1}{N} \sum_{n=0}^{N-1} F_m e^{\left(\frac{-i2\pi mn}{N}\right)} \tag{1.48}$$



(a)



(b)



(c)

Figure 1.10: Demonstration of the rectangular approximation to the continuous Fourier integral employed by the DFT. The determination of the real part of the $n = 1$ Fourier coefficient is shown.
(a) The rectangular approximation to $f(x)$.
(b) The rectangular approximation to $\cos(\frac{2\pi}{N})$.
(c) Evaluation of the Fourier coefficient which is equal to the sum of the areas as shown.

Whilst (1.47) can be invoked to generate the Fourier spectrum on a sampled grid, it is often necessary to generate the continuous periodic spectrum which corresponds to a sampled image. By allowing the index m to be continuous it is possible to interpolate between the samples F_m. The extent W of the continuous spectrum $F(u)$ can then be related to the extent X, of the image and the number N, of sample points, by

$$W = \frac{N}{2X} \tag{1.49}$$

It is therefore apparent that knowledge of two out of N, W, and X is sufficient to completely determine the relationship between the DFT and the continuous Fourier transform. Using (1.47) and (1.48) to implement the DFT is inefficient as the computational complexity grows proportionately to N^2. It was not until the advent of a new algorithmic implementation of the DFT, namely the fast Fourier transform (or FFT) by Cooley and Tukey (1965), that its use became widespread (Bates and McDonnell 1986 §12). The algorithm has in the one-dimensional case a complexity proportional to $N \log_2 N$, thereby providing significant improvements in both speed and round-off noise.

It is apparent that the DFT approximates the continuous Fourier transform by approximating the image function with a series of rectangles, as shown in Fig 1.10. Although this may appear a fairly crude approximation to the continuous Fourier integral, it is well known that there is always a tradeoff to be made between the rate of sampling and the sophistication of the interpolation between samples (Kreysig 1979, p784). In the case of the DFT, the speed of calculating the FFT more than compensates for the simplistic approximation to the continuous integral.

Chapter 2

APPLICATIONS

The process of convolution is described in §1.2. The reverse process, deconvolution, has been for many years the subject of widespread interest. Although in many ways theoretically straightforward, deconvolution is in fact fraught with practical difficulties (Cornell 1987). In this thesis the following notation is used to model a practical convolution

$$g(\vec{x}) = f(\vec{x}) \odot h(\vec{x}) + c(\vec{x}) \tag{2.1}$$

or equivalently, as required by the convolution theorem (1.24),

$$G(\vec{u}) = F(\vec{u})H(\vec{u}) + C(\vec{u}) \tag{2.2}$$

where $f(\vec{x})$ is the true (or undistorted) image, $h(\vec{x})$ the psf or blurring and $c(\vec{x})$ accounts for measurement noise and also model deficiencies (e.g. aliasing as covered in §1.5, or nonlinearity as covered in §2.2). The true image can only be recovered exactly in the unlikely event that $c(\vec{x})$ is also known exactly. This chapter introduces a series of practical situations which require deconvolution, as well as techniques which can be used to achieve this end.

Wiener (1942) originated the most well known form of deconvolution, a process which is now known as Wiener filtering. Wiener filtering requires an estimate of $h(\vec{x})$ (this is in contrast to the new methods of deconvolution discussed in chapters 4 and 6). Despite having been the subject of intensive research, the practical application of Wiener filtering is still an open area (Bates and McDonnell 1986). §2.1 describes Wiener filtering and how it can be used to correct motion blurring in photography.

The astronomical setting is introduced in §2.2. Although most people are familiar with the twinkling of stars, the exact nature of this distortion is less well known and is discussed in §2.2. The majority of this section is devoted to how the atmosphere can be modelled. Atmospheric modelling is of prime importance because the detail in astronomical photographs is usually determined less by the properties of the instrument used than by the quality of the seeing (or atmospheric conditions). §2.2 shows how atmospheric blurring causes a reduction in the higher spatial frequency content and a consequent loss of sharpness in the image.

One method for overcoming the loss of higher spatial frequencies caused by atmospheric turbulence is to limit the exposure time. If the time of exposure is sufficiently short then the turbulent atmosphere is effectively rendered stable. If a short exposure image is restricted to quasi-monochromatic light, as characterised by (1.16), a speckle image results. A speckle image has a mottled appearance reminiscent of a bunch of

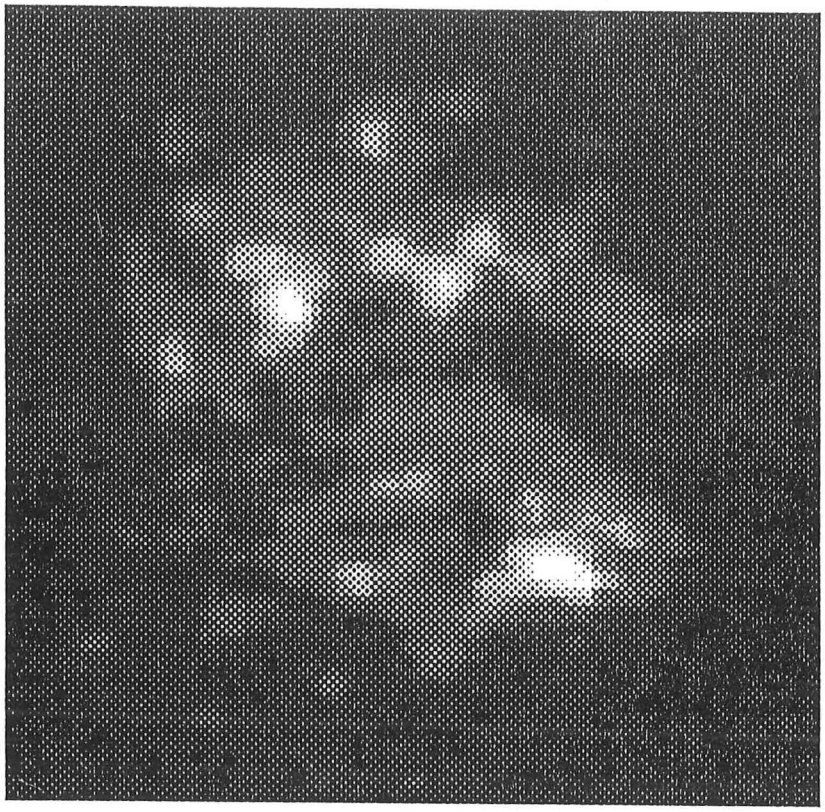


Figure 2.1: Speckle image of Betelgeuse, formed in 4m telescope. Light frequency = 600 nm, light bandwidth = 2nm. Figure is quantised into 32 grey levels, from 0 (black) to a normalised maximum of 1 (white).

grapes (see Fig 2.1). The speckled appearance results from phase distortion of higher spatial frequencies (Roddier 1981). The visibility magnitude is, however, usually less distorted. §2.3 describes Labeyrie's speckle interferometric processing which uses sequences of speckle images to estimate the visibility magnitude (Labeyrie 1970).

As is noted in the Preface, the visibility magnitude requires an estimate of the visibility phase before it is possible to reconstruct an estimate of the true image. This and other aspects of recovering the Fourier phase are discussed in §2.4, which also defines what is here called the Fourier phase problem (Bates and Mnyama 1986).

The converse problem, that of recovering a visibility's magnitude from its phase is addressed in §2.5. The magnitude problem, as it is referred to in this thesis, has been much studied theoretically but it has had little practical application. In this thesis, however, it forms an important step in a new method of blind (i.e. not requiring prior knowledge of the detailed form of the psf) deconvolution which is the topic of chapter 6.

The chapter concludes in §2.6 with a brief discussion of other forms of blind deconvolution, namely shift-and-add, homomorphic filtering and maximum entropy. This section is not intended to present a comprehensive analysis, but to provide a background of alternatives to the methods described in this thesis.

It should be noted that nearly all deconvolution algorithms rely on some form of subjective input. In practical methods of deconvolution the deconvolved image is not sensitive to the assumptions made. Thus the various methods of deconvolution are to a large degree complementary.

2.1 Wiener Filtering

The technique of Wiener filtering is a development of what has been called simple inverse filtering. If in (2.2) the contamination $C(\vec{u})$ is identically equal to zero and if $H(\vec{u})$ is known, it is then possible to reconstruct $f(\vec{x})$ using

$$f(\vec{x}) \longleftrightarrow \frac{G(\vec{u})}{H(\vec{u})} = \frac{F(\vec{u})H(\vec{u})}{H(\vec{u})} = F(\vec{u}) \tag{2.3}$$

In practice, however, $C(\vec{u})$ is not zero and simple inverse filtering leads to an estimate $f'(\vec{x})$ of the true image given by

$$f'(\vec{x}) \longleftrightarrow \frac{G(\vec{u})}{H(\vec{u})} = F(\vec{u}) + \frac{C(\vec{u})}{H(\vec{u})} \tag{2.4}$$

This estimate is deficient because whenever $H(\vec{u})$ is small the effect of the noise $C(\vec{u})$ is amplified. Simple inverse filter estimates are usually characterised by "ringing" in the image at frequencies corresponding to where $H(\vec{u})$ is small (compared with the contamination). In order to avoid this distortion it is customary to employ the Wiener filter estimate

$$f'(\vec{x}) \longleftrightarrow G(\vec{u})\frac{H^*(\vec{u})}{|H(\vec{u})|^2 + |\Phi(\vec{u})|^2} \tag{2.5}$$

where $|\Phi(\vec{u})|^2$ is an appropriately normalised estimate of the power spectrum of the noise. Although the Wiener filter can be shown to be least squares optimal over an ensemble of statistically similar images (Andrews and Hunt 1977), it is not necessarily optimal for any specific image. In addition detailed estimates of the noise statistics are seldom available and it is thus necessary to estimate $|\Phi(\vec{u})|^2$. One way of estimating $|\Phi(\vec{u})|^2$ is simply to assume it to be equal to a real positive constant Φ_0, as would be the case if $c(\vec{x})$ was additive white noise (Bates and McDonnell 1986).

The frequent necessity of estimating $|\Phi(\vec{u})|^2$ leads to a more pragmatic assessment of Wiener filtering than those based on the formation of least squares optimal estimates. Bates (1982a), Bates et al (1984), McDonnell (1975), all emphasise the role $|\Phi(\vec{u})|^2$ in enhancing those spatial frequencies having a high signal to noise ratio.

Fig 2.2 shows a computer simulation of motion blur which occurs when a picture is taken by a moving camera. Each point in the undistorted image (i.e. the image that would have been taken had the camera not been moving) is spread onto a line segment (an image which McDonnell (1975) calls linearly blurred), as shown in Fig 2.2b. The magnitude of the psf visibility and the blurred image are shown in Figs 2.2c and 2.2d respectively.

The reconstructions obtained from the simple inverse and Wiener filter estimates are shown in Figs 2.3a and 2.3b respectively. In this instance the simple inverse filter estimate bears very little resemblance to the true image. The Wiener filter, however, to a large degree eliminates the ringing inherent in the simple inverse filter estimate (although there is still some distortion apparent in Fig 2.3b).

Another difficulty that often occurs in the astronomical context is that the image formed is too large for the instrument, or in other words the instrument's field of view is less than the angular field spanned by the atmospherically blurred object. Setting to zero those parts of the image which are truncated can also significantly affect the reconstruction (McDonnell 1975). This distortion can to a large degree be mitigated by either windowing or edge-extension (Bates and McDonnell 1986).

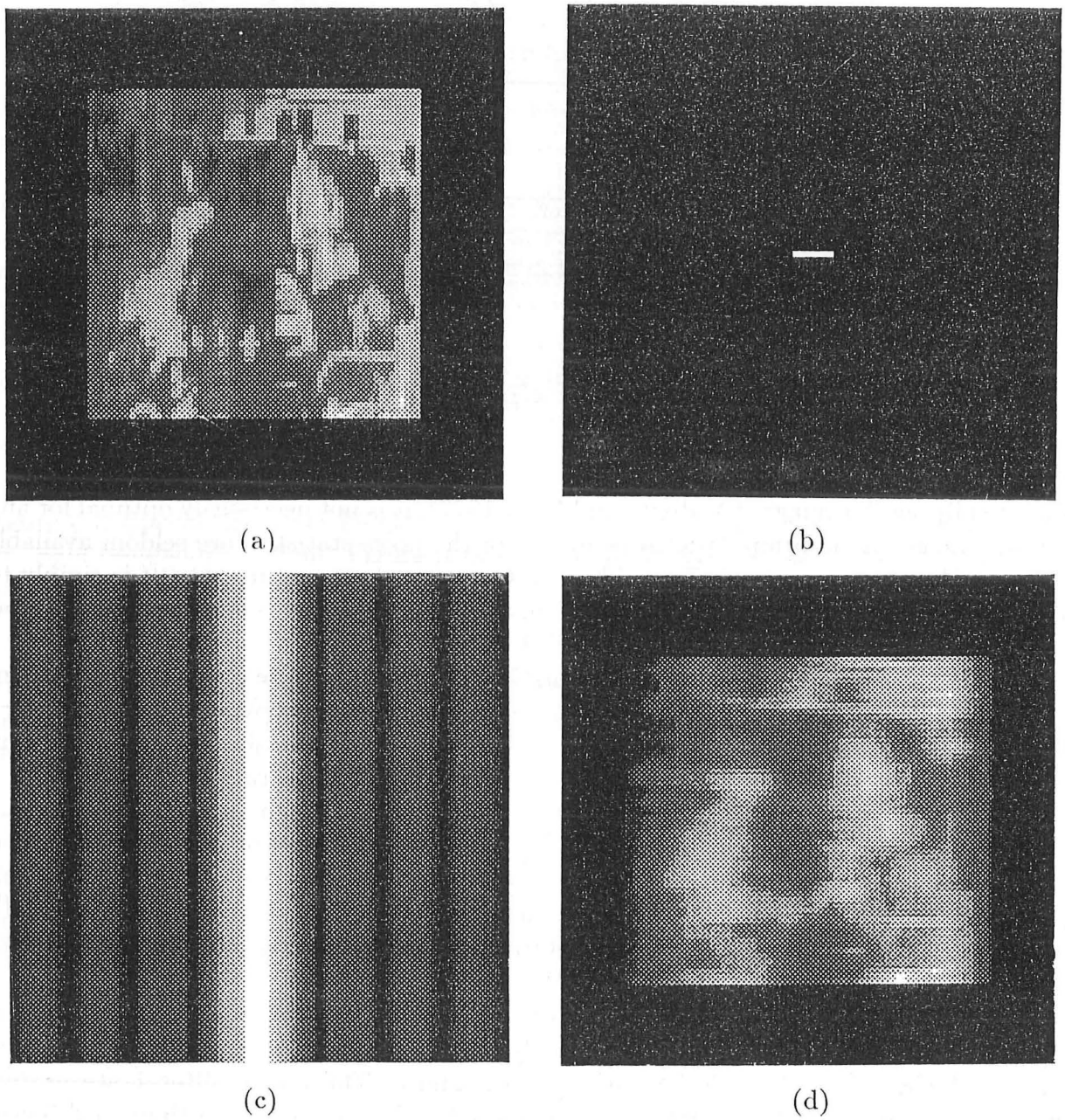
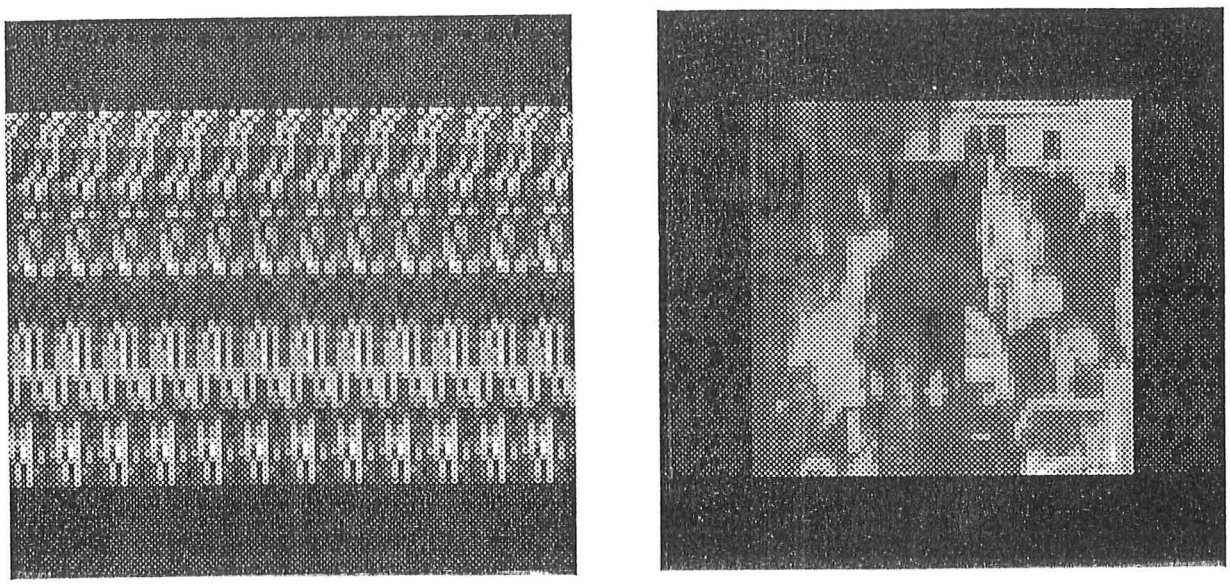


Figure 2.2: Illustration of motion blur. Quantised as in Fig 2.1. (a) true image $f(x, y)$, (b) psf $h(x, y)$, (c) psf visibility $H(u, v)$, (d) linearly blurred image $g(x, y)$



(a) (b)

Figure 2.3: Deblurred estimates of the true image shown in Fig 2.2a. Quantised as in Fig 2.1. (a) Inverse filtered estimate from noisy version of Fig 2.2b, (b) Wiener filtered estimate from noisy version of Fig 2.2b

2.2 The astronomical setting

Astronomy is perhaps one of the more difficult of the physical sciences because it is not usually possible to actively experiment with the objects of interest. These objects can be divided into two classes, passive radiators such as the moon and planets, and active radiators such as stars. In either case the sources of wave motion are almost always spatially incoherent (cf §1.2) and sufficiently distant that a Fourier transform exists between the object and the radiation pattern incident on the earth. This is complicated, however, by the existence of a turbulent atmosphere between earthbound observers and the celestial objects they wish to observe. The wave motion is distorted by variations in the atmospheric refractive index caused by temperature fluctuations (Roddier 1981). This distortion is manifest as the familiar twinkling of stars when observed with the naked eye. A more detailed analysis of the "seeing problem" can be found in the following works and their references: Sinton (1986), Bates (1982a), Dainty (1982), Roddier (1981).

Consider the idealised instrument shown in Fig 2.4, which is forming an image of a distant stellar object. Due to the vast interstellar distances the incoming radiation is effectively a set of plane waves. The telescope transforms the angle of the incoming plane waves onto a position on the focal plane of the object (Born and Wolf 1970, §7.3.6). This process effectively inverts the Fourier transform relationship between an object and its far-field radiation pattern (Born and Wolf 1970, §8.3.3, this thesis §1.2). Since the telescope aperture (or pupil) is of limited size only a small portion of the visibility spectrum contributes to the observed intensity at any particular point in the image. This results in a loss of image detail because the part of the visibility spectrum lost corresponds to the higher spatial frequencies of an object. The image thus formed is termed diffraction

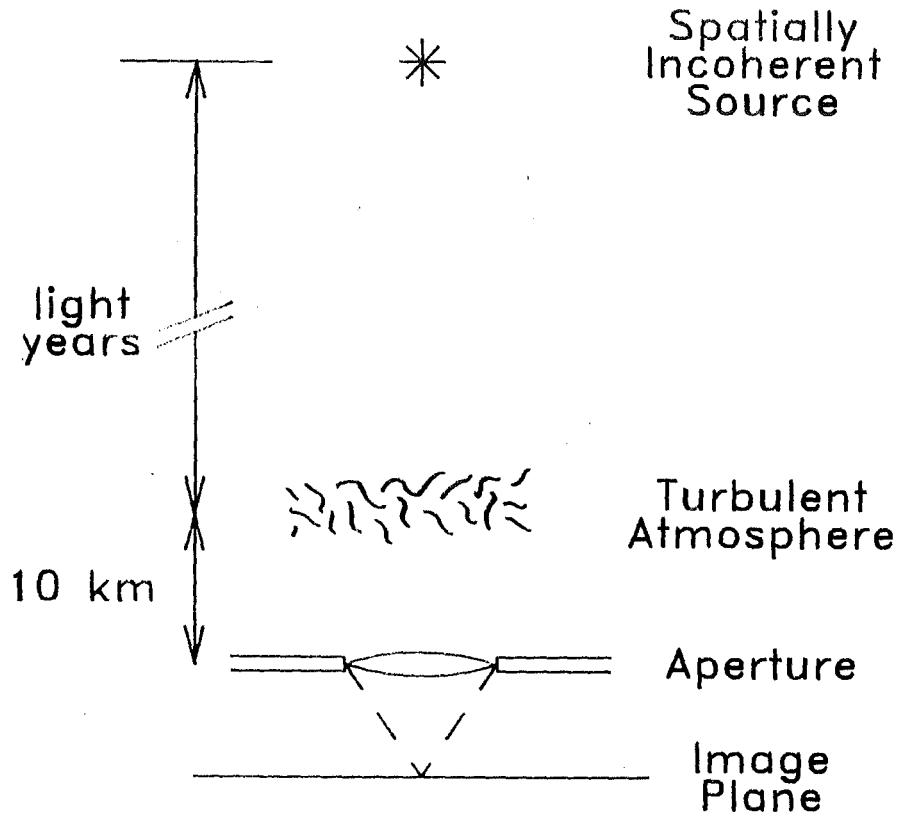


Figure 2.4: Diagramatic representation of the astronomical setting

limited.

The truncation of the Fourier spectrum can be modelled by a suitable low-pass filter $A(u,v)$, for example Fig 2.5a. The image spectrum is of the form

$$F(u,v) = O(u,v)A(u,v) \tag{2.6}$$

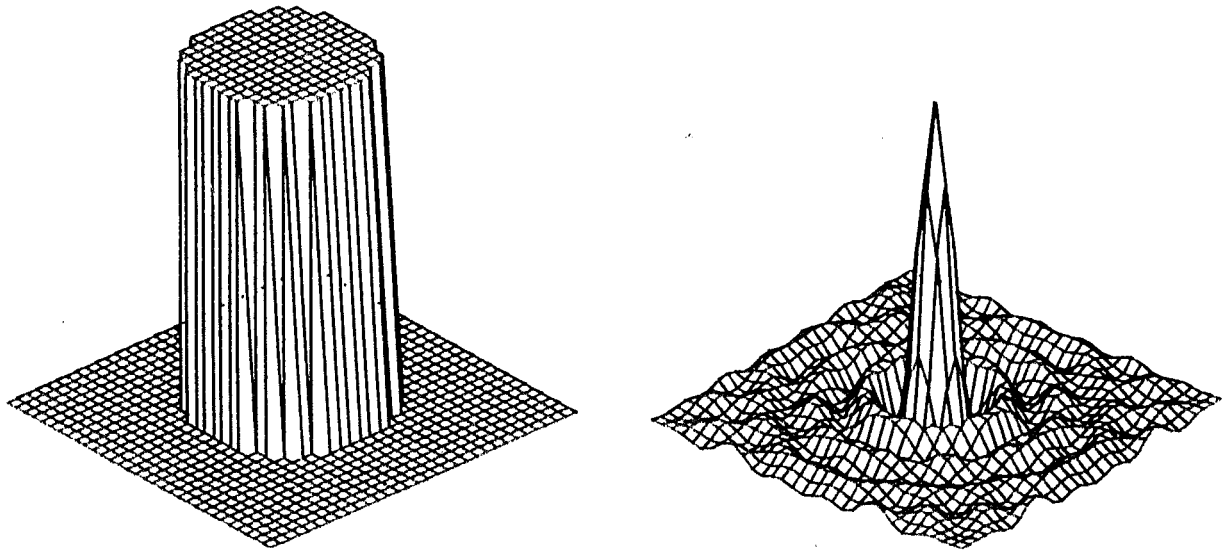
where $O(u,v)$ is the spectrum of the equivalent source of the object (§1.2) and $A(u,v)$ models how only a restricted portion of the object's far-field radiation pattern is observed in an instrument of finite dimension. Hence, the observed image $f(x,y)$ is the convolution of the object's equivalent source (cf §1.2) and the aperture function $a(x,y)$.

$$f(x,y) = o(x,y) \odot a(x,y) \tag{2.7}$$

It is apparent from (2.6) that $a(x,y)$ models the response of the telescope to a unresolvable object (Fig 2.5b) since $f(x,y)$ can be modelled by a delta function.

One critical measure of a telescope's performance is its resolution, or its ability to separate closely spaced objects. As $A(u,v)$ represents the aperture of the telescope, it can only pass the lower visibility frequencies and the higher frequencies present in $O(u,v)$ are lost. Furthermore if $A(u,v)$ has a sharp cutoff in the frequency domain this can introduce undesirable ringing in $f(x,y)$, because $a(x,y)$ inevitably has large sidelobes (Fig 2.5b). These sidelobes can be reduced by deliberately tapering $A(u,v)$ at higher frequencies, a process known as apodisation (Born and Wolf 1970, p 417).

For a simple circular aperture of radius D, the spatial frequency limit is given approximately by $\frac{D}{\lambda}$. A point source observed through a circular aperture forms a diffraction limited image known as the Airy pattern. The central lobe of this pattern is known as the Airy disc and is approximately $\frac{\lambda}{D}$ in diameter. Because resolution decreases with



(a) (b)

Figure 2.5: The aperture function of a diffraction limited instrument, (a) $A(u,v)$ (b) $a(x,y)$.

increasing wavelength it is necessary to build instruments with larger apertures if one wishes to obtain a similar spatial resolution at longer wavelengths.

The aperture size provides a fundamental limit on the resolution for a given instrument. In the case of visible light and large optical telescopes, however, the atmosphere is a more serious cause of image distortion. Utilising the notation and terminology of §1.2, the atmospheric distortion is described by $h(x_e, y_e, x_o, y_o, t)$, a function of position in both the image and object planes, as well as of time. The time variation of $h(x_e, y_e, x_o, y_o, t)$ is usually summarised in the form of two limiting intervals. The first, t_s, is the time for which

$$h(x_e, y_e, x_o, y_o, t + t_s) \approx h(x_e, y_e, x_o, y_o, t) \tag{2.8}$$

and defines the time for which the atmosphere is effectively stationary. The second time constant, the redistribution time t_r, models the time taken for $h(x_e, y_e, x_o, y_o, t)$ and $h(x_e, y_e, x_o, y_o, t + t_r)$ to become uncorrelated.

A useful simplification of the point spread function is obtained by modelling the atmosphere as an amalgamation of blobs, or seeing cells. The approximate size of these cells is given by the parameter d_b. The time delay experienced by radiation passing through a particular seeing cell alters the phase of the radiation incident on the observation plane. Whilst at a particular time the observed phase of radiation passing through a given seeing cell may be assumed to be constant, it is random with respect to the phase of radiation passing through other seeing cells. Since the atmosphere is very much closer to the telescope than the object, nearly all the light incident on any particular point in the image plane passes through a single seeing cell (Fig 2.6). The blurring is then isoplanatic, and can be described by $h(x_o, y_o, t)$ (henceforth written as $h(x, y, t)$). The validity of this approximation is to a large degree a function of the size of the isoplanatic patch (Sinton 1986), but is often a good model when small deviations from isoplanitism

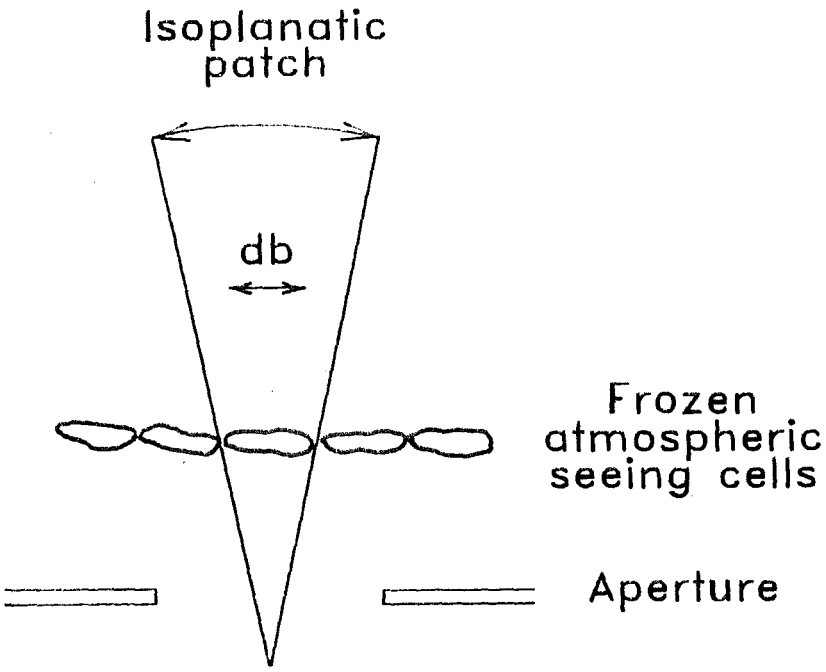


Figure 2.6: Approximation of the atmosphere by a collection of seeing cells

are accomodated as an additive noise ($C(\vec{u})$ in (2.2)).

In many astronomical photographs the object of interest exhibits apparent motion due to the earth's rotation. This causes difficulty with faint astronomical objects because a long exposure time is required in order to gather sufficient light to form an image. In long exposure images it is necessary to compensate for the object's apparent motion by moving the telescope appropriately. The image obtained by tracking the object is thus a time average

$$\overline{s(x,y,t)} = \overline{f(x,y,t)} \odot \overline{h(x,y,t)} + \overline{c(x,y,t)} \tag{2.9}$$

where the superscript bar denotes a time average.

The major difference between $\overline{h(x,y,t}$ and $h(x,y,t)$ is best discussed in terms of their Fourier transforms. In $H(u,v,t)$ the spatial frequencies above $\frac{d_h}{\lambda}$ pass through different seeing cells and their phases are effectively randomised. When a time average is performed these spatial frequencies tend to cancel, causing $\overline{|H(u,v,t)|}$ to fall off very rapidly once $|(u,v)|$ exceeds $\frac{d_h}{\lambda}$. Loss of these higher spatial frequencies causes a point source to be smeared into the seeing disc, of width sd. Thus for a 5m telescope, the resolution limit imposed by the atmosphere can typically be of the order of 50 times less than the potential (i.e. diffraction limited) resolution of the telescope.

The form of $\overline{H(u,v,t)}$ also presents significant difficulties to the process of deconvolution, even if it can be estimated accurately. In the discussion of the Wiener filter (§2.1) it is noted that when the signal is small in comparison to the noise level, the Wiener filter simply serves to attenuate these frequencies. In practical terms, when spatial frequencies have been reduced below the level of the noise they are irrecoverable.

Table 2.1 (after Roddier 1981) gives typical values of the parameters used in describing atmospheric distortion. The quoted values can vary, either way, by an order

Parameter	Symbol	Typical value
blob size	d_b	100 mm
time atmosphere effectively stationary	t_s	10 ms
redistribution time of atmosphere	t_r	500 ms
isoplanatic patch size		10 arc seconds
seeing disk size	sd $\frac{\lambda}{d_b}$	1 arc seconds
Airy disk size (5m telescope)	Ad $\frac{\lambda}{D}$	0.025 arc seconds

Table 2.1: Typical values of atmospheric distortion for 500 nm light

of magnitude.

2.3 Speckle

In order to preserve the spatial frequencies for which $|(u,v)|$ is greater than $\frac{d_b}{\lambda}$, Labeyrie (1970) proposed that the image be exposed for a short time no longer than t_s. The higher spatial frequencies are thus preserved, although their phases are disturbed. This phase error in the higher spatial frequencies causes the light in the short term exposure image to be distributed over the seeing disc. Unlike the smooth long term exposure image the short term exposure image assumes a granular appearance. This short term exposure image is made up of a number of speckles, each of which resemble in some way the original object (Bates 1982a). The high frequency detail is thus preserved, albeit in a confused form, in the short exposure image (or speckle pattern). The number of speckles is a function of the ratio of the size of the seeing disc to the Airy disc.

Speckle imaging can, in practice, be successful only if a large number of speckle images are recorded. Ideally, different speckle patterns should be separated in time by t_r, so that they are statistically independent. In practice, this criterion is rarely met and the time between each speckle pattern is determined by the speed of the recording equipment. As noted by Bates (1982a) the effective number of independent images is given approximately by the observation time divided by the redistribution time of the atmosphere.

Once the data has been collected there remains a number of linked deconvolution problems,

$$s_m(x,y) = f(x,y) \odot h_m(x,y) + c_m(x,y) \text{ for } m = 1, \cdots, M \tag{2.10}$$

which can be Fourier transformed to yield

$$S_m(u,v) = F(u,v) \odot H_m(u,v) + C_m(u,v) \text{ for } m = 1, \cdots, M \tag{2.11}$$

where $H_m(u,v)$ is the m^{th} optical transfer function (OTF). When these are simply averaged the result is the same as for a long term exposure image, the higher spatial frequencies

simply cancel. Labeyrie (1970) proposed that only the intensity of the Fourier transform be averaged over the ensemble (or collection of speckle images). Using <> to denote the process of averaging over the ensemble:

$$< |S_m(u,v)|^2 > = |F(u,v)|^2 < |H_m(u,v)|^2 > + < E_m(u,v) > \qquad (2.12)$$

where $< E_m(u,v) >$ is a real term containing all the cross products which include the contamination. The next stage is to estimate $< |H_m(u,v)|^2 >$. This is done by pointing the telescope at an unresolvable star and repeating the above process under statistically similar seeing conditions (ideally during the observation of the object of interest):

$$< |S^0_m(u,v)|^2 > = |F^0(u,v)|^2 < |H^0_m(u,v)|^2 > + < E^0_m(u,v) > \qquad (2.13)$$

which is a useful estimate of the unknown $< |H^0_m(u,v)|^2 >$, because $|F^0(u,v)|^2$ is effectively constant (i.e. it does not vary with u and v, by definition, since the star is said to be unresolvable). Thus, on account of the stated similarity of the seeing conditions, $< |H_m(u,v)|^2 >$ can be taken to be directly proportional to $< |S^0_m(u,v)|^2 >$. By employing Wiener filtering it is then possible to estimate $|F(\vec{u})|^2$. The use of an unresolvable reference star has the added advantage that the effects of the aperture, and to some extent the telescope imperfections, are also corrected (Dainty 1973; Bates and Cady 1980).

The above description is an outline of a simple speckle imaging technique which produces an estimate of the Fourier magnitude of the object. There are alternative methods of estimating the Fourier magnitude (Bruck and Sodin 1984) and correcting for various types of noise such as photon noise (Feldkamp and Fienup 1980). Labeyrie's processing provides a pure example of the so called "phase problem" (see §2.4).

Although techniques have been developed for estimating the Fourier phase from ensembles of speckle images (Bates 1982a, §8), it is important to realise that the magnitude and phase of an object are measured to the same accuracy only rarely. As a consequence there is always a place for algorithms which can use interrelationships between magnitude and phase to improve poor measurements of one or the other.

2.4 The phase problem

The previous section describes a situation where it is possible to measure the magnitude but not the phase of an image's visibility. There are in fact a number of reasons why the visibility phase may be difficult to measure accurately:

1. Accurate receivers sensitive to the phase of high frequency measurements may be expensive or difficult to obtain (Morris 1985).

2. The phase of the wave motion may be distorted during propagation, as is the case with atmospheric distortion (Bates 1982a).

3. The instrument may have inherent inaccuracies which cause a distortion in the measured phase, e.g. misalignment of antenna panels (Morris 1985).

4. The phase may be discarded to increase storage and/or transmission efficiency as is the case with LPC coding of speech signals (Makhoul 1975).

5. There may be difficulties in maintaining stable phase references when measurements are made at widely separated geographic sites, as occurs in Very Long Baseline Interferometry (Readhead et al 1980).

In most situations, however, the phase is not totally lost and some estimate of it can be made. Although rarer and more difficult, the pure phase problem (i.e. situations where the phase is totally lost) provides useful insight into the interrelationship between Fourier magnitude and phase. This thesis only considers solving the pure phase problem because any techniques capable of solving the pure phase problem can usefully be applied to partial phase problems.

In addition to the above restriction only phase problems where it is possible to form the autocorrelation of the object via (1.27) are considered. This constitutes the Fourier phase problem (Bates and Mnyama 1986), and is equivalent to requiring that the Fourier magnitude be oversampled by a factor of two (§1.5). In situations where the Fourier transform is continuous this is not difficult and the algorithms in this thesis have application in a wide range of fields (Table 2.2). A notable exception is the crystallographic phase problem where the visibility is in itself discrete at the Nyquist frequency. The inverse Fourier transform of the Nyquist sampled magnitude yields an aliased version of the autocorrelation known as the Patterson (Ramachanran and Srinivasan 1970).

Application	Causes	Type
X-ray and neutron crystallography	wavelength too short for sensing	pure but non Fourier
electron microscopy	wavelength too short for sensing instrument imperfections	partial
radio astronomical aperture synthesis	turbulent medium instrument instability	partial
optical astronomical speckle interferometry	turbulent medium	pure
optical astronomical speckle imaging	turbulent medium	partial
radio engineering	accuracy depends on wavelength	partial
ultrasonics acoustics	distorting medium	partial
communications speech processing	economy of storage	pure and partial

Table 2.2: Practical occurrence of the phase problem

Recovery of the crystallographic phase to a large extent relies on the knowledge that the crystal structure is comprised of a number of atoms. If this were not so the problem would be insoluble (Karle 1986; Hauptmann 1986). In the crystallographic phase problem significant a priori information can be obtained by chemical procedures such as

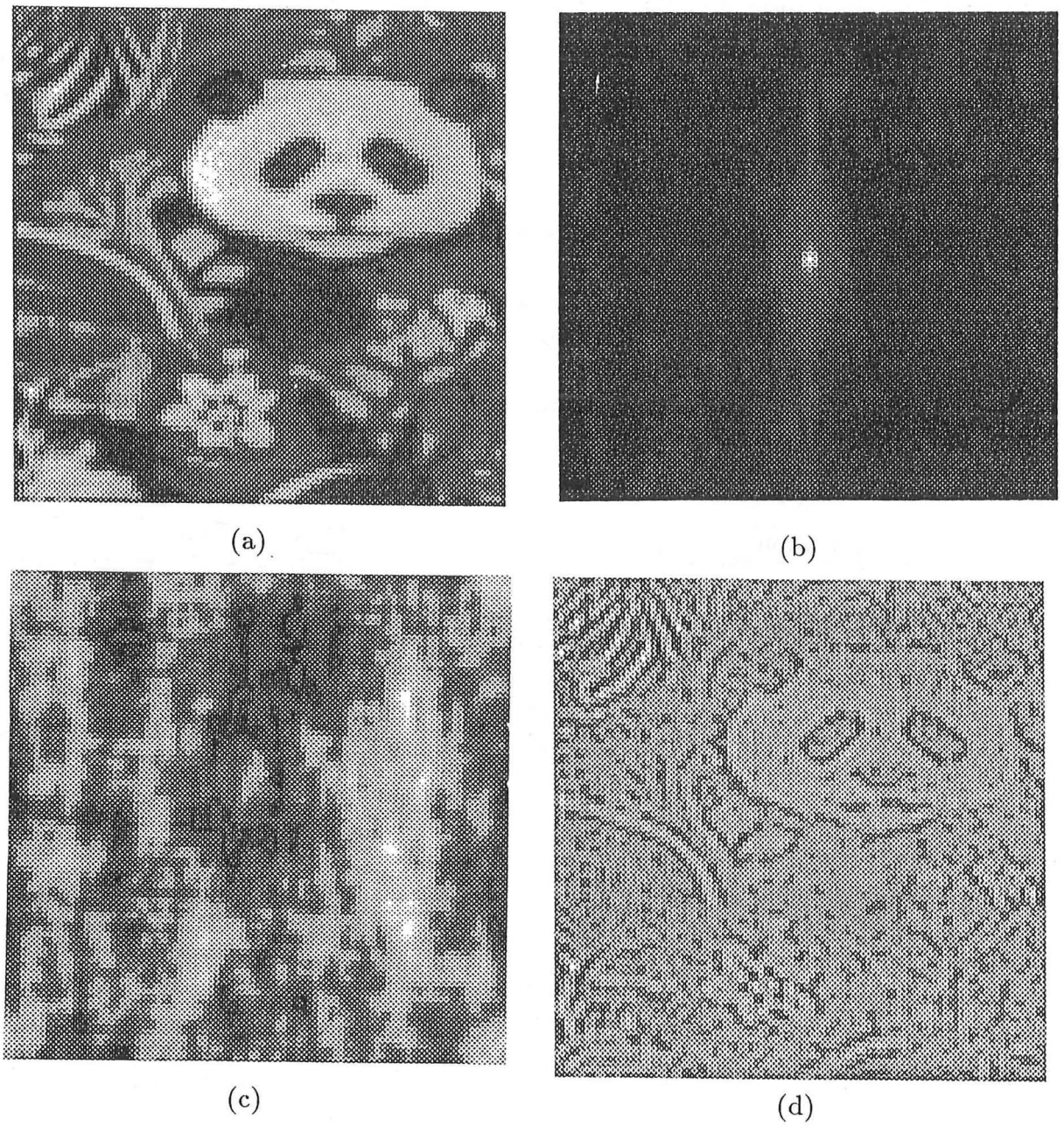


Figure 2.7: Estimates of an image formed using partial information in the Fourier domain. Figures are displayed in 32 grey levels from most negative (black) to most positive (white). (a) true image (b) true visibility magnitude with zero phase (c) true visibility magnitude with random phase (d) constant visibility magnitude with true phase.

isomorphous replacement. Thus crystallographic phase retrieval, although interesting, falls outside the realm of this thesis.

Just how important the phase is can be seen from Fig 2.7. Combining the visibility magnitude with either zero or random phase produces an image bearing no resemblance to the original object.

2.5 The magnitude problem

The magnitude problem (i.e. the recovery of a visibility's magnitude from its phase) has also been the subject of intensive theoretical investigation (Anderson and Anderson 1986; Bruck and Sodin 1983; Hayes 1982; Oppenheim and Lim 1981; Kermisch 1970). There has also been some practical investigation of its application to synthetic aperture radar (Munson and Sanz 1986) and speckle techniques (Bruck and Sodin 1984). The magnitude problem has been thought to be easier than the phase problem for a number of reasons. Firstly it is possible in theory to solve the magnitude problem by using linear systems of equations. This procedure has been demonstrated by Bruck and Sodin (1983). Secondly the visibility of most common objects is lowpass in nature. Replacing the Fourier magnitude with a uniform magnitude is thus equivalent to high-pass filtering the object, a process known to enhance the edges in the object. As the edges play an extremely important function in how the human visual system (Marr et al. 1979) interprets an image, the image is in general still easily recognisable, as shown in Figs 2.7d. The quality of the reconstruction is improved further if an approximate estimate of the Fourier magnitude is made (Oppenheim and Lim 1981).

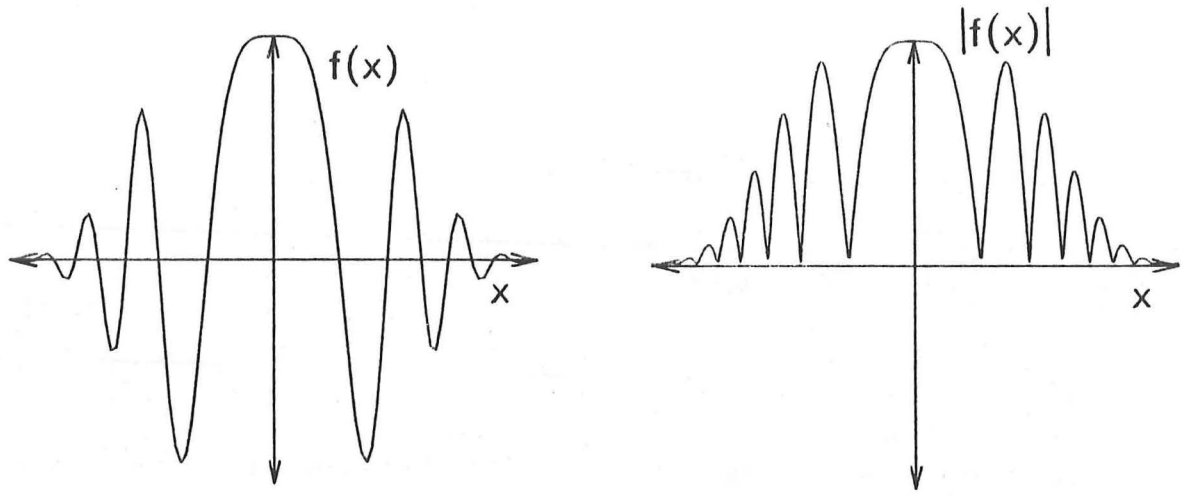
When displaying phase-only reconstructions it is important to realise that the DC level is lost and it is therefore necessary to add an offset to make the image positive before it is displayed. Taking the magnitude of a bipolar (i.e. positive and negative) image results in an output image which is discontinuous at the zero-crossings of the input image, as illustrated in Fig 2.8. Comparison of Figs 2.7d and 2.9 shows the dramatic difference in the appearance betweeen the magnitude of a bipolar image (Fig 2.9) and the image obtained by adding a positive offset (Fig 2.7d). Whilst many authors are aware of this problem (e.g. Hayes 1982, Oppenheim and Lim 1981), it is possible that this may have been overlooked by other authors (e.g. the images presented by Haque and Meyer 1986 appear to suffer this form of distortion).

The magnitude problem finds application in a number of areas, e.g. phase only holograms or kinoforms (Lesem et al 1969). Reconstructions obtained by assuming that the unknown magnitude is constant have often proved perfectly adequate (Hayes 1982, Oppenheim and Lim 1981, Hayes et al 1980). A more difficult problem, and one which finds application in this thesis, is where the phase is only known modulo π. This arises, for example, when the true image $f(\vec{x})$ is convolved with a symmetric point spread function $s(\vec{x})$ to form a convolution $g(\vec{x})$. Since the visibility of a symmetric image is real (Bates and McDonnell 1986; Bracewell 1978) it follows that

$$\mathcal{P}[F(\vec{u})] = \mathcal{P}[G(\vec{u})] \quad or \quad \mathcal{P}[G(\vec{u})] \pm \pi \tag{2.14}$$

Because the phases are computed, as opposed to analytically determined, $\mathcal{P}[F(\vec{u})]$ can be computed modulo π directly from the phase of $G(\vec{u})$.

Unlike the straightforward magnitude problem, reconstruction from the phase modulo π presents significant difficulties. Fig 2.10 shows an estimate formed by combining



(a) (b)

Figure 2.8: Difference between a bipolar one-dimensional image, and its magnitude. Note whenever the true image crosses zero there is a discontinuity in the slope of the magnitude function, (a) true bipolar image (b) magnitude of bipolar image.

Figure 2.9: Magnitude of the image shown in Fig 2.7d. Quantised as in Fig 2.1.

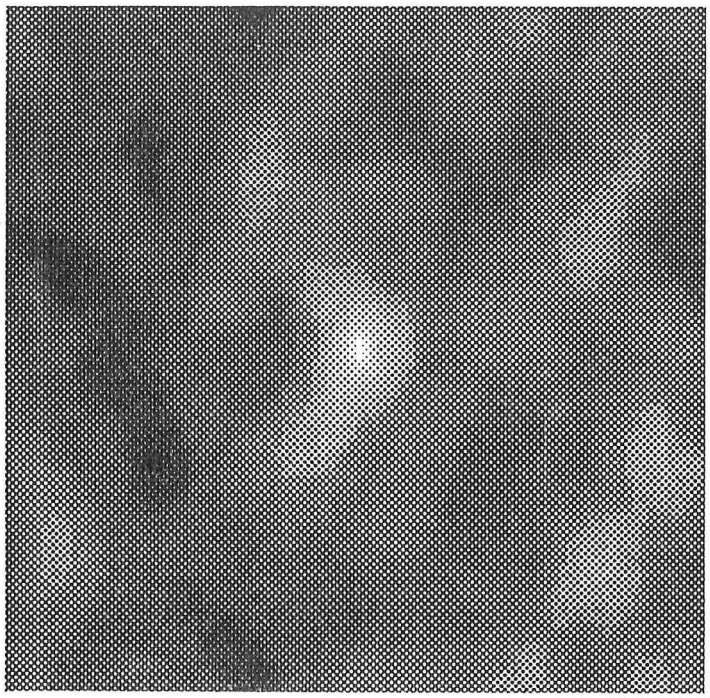


Figure 2.10: Estimate of the true image shown in Fig 2.7a formed by combining a constant magnitude with the true phase modulo π

a uniform magnitude with the phase modulo π. §6.2 discusses the solution of this problem and introduces a technique for deriving the modulo π phase from a general convolution.

2.6 Methods for blind deconvolution

It is probably fair to state that, in the opinion of most image processing experts, it is necessary to possess an accurate estimate of the psf in order to carry out deconvolution successfully. Furthermore, it is currently assumed that it is necessary to have an ensemble of differently (but statistically similarly) blurred images of an object before an image can be generated by "blind deconvolution" (i.e. image recovery in the absence of a priori information about the psf).

By contrast the methods presented later in this thesis (chapters 4 and 6) can, in theory and to some extent in practice, effect deconvolution from a single convolution. The purpose of this section is to introduce the techniques based on deconvolving ensembles of similarly blurred images, in order to provide a comparison with the algorithms presented later in this thesis.

The primary assumption made when deconvolving from an ensemble of similarly blurred images is that whilst one component of the convolution is constant, the other is in some sense random. The processing then tries to combine these individual convolutions so as to "average out" the effects of the randomly varying component, for example the atmospheric distortion in speckle imaging (§2.2).

2.6.1 Shift-and-add (SAA)

As well as the Fourier techniques for speckle imaging discussed in §2.3, it is possible to effect blind deconvolution by a process known as shift-and-add (Bates and Cady 1980). This algorithm relies on shifting the brightest point of each speckle image to the centre of image space before performing the ensemble average. As is noted in §2.2, a speckle image has the appearance of the true image randomly shifted and superimposed many times. Thus if the true image has a single dominant bright spot, as is often the case in astronomy, the shift-and-add image consists of the true image on a background fog. The technique has a pleasing simplicity and can be extended for cases where the image has several brightest points (Davey et al 1986; Sinton et al 1986).

2.6.2 Homomorphic deconvolution

An alternative method is based on homomorphic deconvolution, a technique pioneered by Oppenheim et al (1968). The technique again relies on having a large ensemble of similarly blurred images, although these may be obtained by partitioning a non-stationary image (Stockham et al 1975). This technique takes the logarithm of the Fourier transform to reduce the convolution (1.24) to the sum of two functions

$$\ln[(G(\vec{u})] = \ln[(F(\vec{u})] + \ln[(H(\vec{u})] \tag{2.15}$$

It is then possible to either subtract $H(\vec{u})$ or remove it by conventional filtering. The estimate is not however totally blind as it requires some estimate of the properties of either $H(\vec{u})$ or $F(\vec{u})$. In the case of deconvolving distortions present in old recordings of the legendary singer Caruso, Stockham et al (1975) estimated the magnitude of the power spectrum of the wanted signal by invoking the power spectrum obtained from a modern singer singing the same work. This enabled an estimate of $|H(u)|$, which proved sufficient to compensate for the magnitude distortion in the original recording.

As noted by Stockham et al (1975) there are significant difficulties in estimating the true phase when using homomorphic deconvolution techniques. Whilst this is unimportant for the processing of monophonic audio signals, because the ear is insensitive to absolute phase (Wang and Lim 1982), §2.5 shows that is crucial for reconstructing two-dimensional images. Stockham et al (1975) only applied homomorphic processing to images distorted by motion blur and defocus.

2.6.3 Maximum entropy method

The final technique mentioned in this chapter, the maximum entropy method, is not so much a method of blind deconvolution (although it finds application in phase retrieval) as a model for the most likely form of an image. Entropy is often useful in regularising a problem where there is insufficient information to guarantee a solution. The concept of entropy has found widespread application in nearly all branches of physical sciences (Frieden 1985, Shannon 1948, Ables 1974, Jaynes 1982, Woodward 1964) and implies a statistical model of image formation.

The exact form of the entropy function is a matter of debate between

$$E_0 = \int \ln[f(\vec{x})]\, d\sigma(\vec{x}) \tag{2.16}$$

and

$$E_1 = \int f(\vec{x}) \ln[f(\vec{x})]\, d\sigma(\vec{x}) \tag{2.17}$$

depending on the basic process assumed (Burg 1978a, Gull and Daniell 1978). The philosophy of MEM is encapsulated in Ables' (1974) "Principle of Data Reduction",

> "The result of any transformation imposed on the experimental data shall incorporate and be consistent with all relevant data and be maximally non-comittal with regard to unavailable data."

By imposing as few constraints as possible on the image, MEM hopes to find the image most likely to have produced the observed measurements.

Although of undoubted utility two points of caution need to be made. Firstly the entropy functional may be extremely complicated, as is undoubtably the case when entropy is used as a constraint in phase retrieval (Narayan 1987). A non-linear functional having multiple maxima can make the process of finding the image having the maximum entropy extremely difficult.

Secondly, the measures of entropy all have the property that

$$\frac{d^2E}{df^2} = f^{-n} \tag{2.18}$$

As noted by Nityananda and Narayan (1982) the above property tends to favour sharply peaked positive images on broad flat backgrounds. Using (2.18) it is possible to construct functionals, e.g. $\int \sqrt{f(\vec{x})} d\sigma(\vec{x})$, which although having no information theoretic background when used as constraints, result in images of comparable quality to those obtained from using (2.16) and (2.17) (Heffernan and Bates 1982).

Whether entropic constraints are always appropriate, especially when dealing with images of highly deterministic (especially man-made) structures is of course open to debate. In the author's opinion MEM is a valid approach capable of producing some good results and is an excellent means of regularising ill-conditioned problems. As with all models, the behaviour of MEM can on occasion be inappropriate, for example in the modelling of ARMA spectral processes (van den Bos 1971). The description of the reconstruction as the most likely image form is also sometimes suspect. Both Burg (1978b) and Frieden (1985) note that in spectral analysis ME estimates are related to, but are not in general equal to, the maximum likelihood estimates.

Chapter 3

ONE DIMENSIONAL MODELLING

Entire functions of a complex variable have found widespread application in the modelling of physical phenomena. In modelling both convolution and phase retrieval many authors (cf Taylor and Whinnery 1951; Bates 1969; Hoenders 1975; Burge et al 1976; Nakajima and Asakura 1983a; Sanz and Huang 1985; Stefanescu 1985) have made use of an entire function model. This is because the Fourier transform of an exactly compact image (§1.4) can be precisely modelled by an entire function.

Although there is a vast and complicated literature concerning the behaviour of entire functions (Boas 1954; Levin 1964; Markusevich 1965; Requicha 1980), much of this is too specialised to find application in the algorithms described in this thesis. Perhaps the simplest description of entire functions is in terms of an infinite Taylor series which is everywhere convergent. As a consequence an entire function has derivatives of all orders and is commonly described as "infinitely smooth". So, if $\Xi(w)$ is such an entire function of the variable w,

$$\Xi(w) = a_0 + a_1 w + a_2 w^2 + \cdots \tag{3.1}$$

In many situations it is possible to approximate $\Xi(w)$ with the first N terms of the above series. This results in a polynomial of order $N-1$, where the order is the highest power of w present.

§3.1 examines the behaviour of finite order polynomials and forms a basis for discussing the more general properties of entire functions. Two forms of polynomial are introduced, namely trigonometric and algebraic, both of which find application in the modelling of the visibility of a sampled image.

It is possible, as noted by Requicha (1980) to accurately model the visibility of an exactly compact image (§1.6), using entire functions of exponential type (or EFETs), and this is the approach that has been taken by most authors. §3.2 shows how EFETs can be used to relate the solutions to the one-dimensional phase-problem. The most important outcome of this means of analysis is the prediction of the existence of ambiguous solutions to the one-dimensional phase problem. It is also possible to relate these solutions by a well defined procedure, colloquially referred to as zero-flipping. §3.2 concludes with an example of this process.

Although §3.2 introduces a means of relating all possible solutions to the one-dimensional Fourier phase problem, there remains, however, the difficulty of obtaining an initial estimate of a phase distribution without the evaluation of infinite products or series. §3.3 describes how the Hilbert transform is used to obtain a particular estimate of

the Fourier phase from the Fourier magnitude. This approach has been investigated by a number of authors, most recently by Nakajima and Asakura (1983a).

The difficulties with the Hilbert transform approach led Napier and Bates (1974) to approximate the visibility by a finite number of its low frequency zeros, corresponding to employing a finite order model of the image. Although it is difficult to prove the theoretical validity of this approach, it is well known that finite order models are perfectly adequate in practice over appropriately restricted intervals in Fourier space. Computational simulation has also left little doubt as to the effectiveness of finite order models. §3.4 discusses the validity of this approximation and compares it with the continuous examples presented in §3.2.

Some interesting properties of the Fourier phase are dealt with in §3.5. These concern the inability to uniquely define the phase of the analytically continued Fourier transform without introducing either a Riemann surface or discontinuities in the Fourier phase. Whilst in one-dimension the visibility phase may in some instances be continuous this is not so in two-dimensions as is discussed in chapter 4.

§3.6 investigates using image positivity as a constraint and shows how the positivity constraint can be rewritten in the form of inequalities, i.e. the phase of samples of the Fourier transform is not determined exactly, but is restricted to a subrange of possible phases. The methods discussed are derived mainly from work originally intended for recovering crystal structures.

The final section of this chapter, §3.7, deals with some of the more subtle differences between polynomials and other kinds of entire functions. Although relevant to phase retrieval these differences are more important when dealing with deconvolution. In fact §3.7 sets the scene for the two-dimensional deconvolutional technique introduced in chapter 4.

3.1 Polynomials

Polynomials are simple effective models of numerous physical phenomena. Signal processing and control theory are examples of fields which make widespread application of the theory of polynomials. An example of a polynomial model is the DFT introduced in §1.7. In order to see the DFT as a polynomial it is convenient to rewrite the forward DFT (1.47) as,

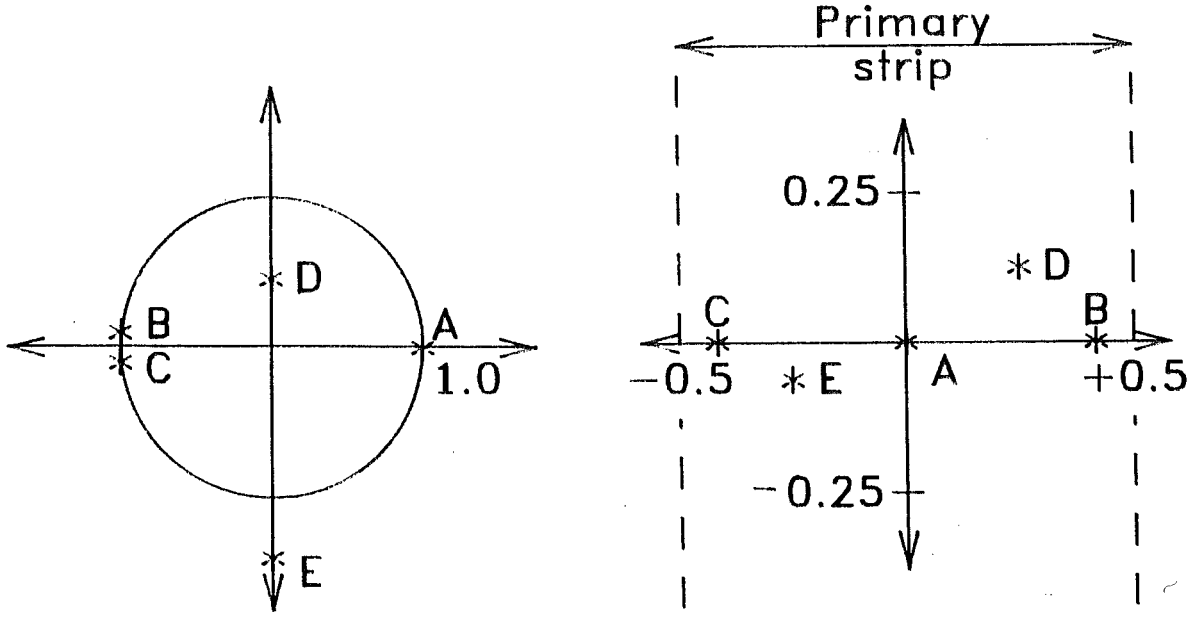
$$F(u) = \sum_{n=0}^{N-1} f_n e^{inu\Omega} \tag{3.2}$$

where Ω is called the fundamental frequency (Requicha 1980) and includes all the scaling necessary to relate the number of samples to the assumed extent in Fourier space (cf §1.7). The f_n are known as the coefficients of the polynomial. The integer N is particularly important, as the quantity $(N-1)$ is known as the order of the polynomial. It is the polynomial order which determines the dimension of the space spanned by the polynomial (Nowinski 1981).

Trigonometric polynomials are periodic functions, i.e.

$$F(\dot{u}) = F\left(\dot{u} + \frac{2\pi n}{\Omega}\right) \quad n = 0, \pm 1, \pm 2 \ldots \tag{3.3}$$

where the period of $F(\dot{u})$ is $\frac{2\pi}{\Omega}$. The dot is placed over the Fourier basis variable u to emphasise that the continuous Fourier variable is being approximated over a finite interval



(a) (b)

Figure 3.1: The conformal mapping between (a) the complex ζ-plane and (b) the complex u-plane ($\Omega = 2\pi$).

by a periodic function. A trigonometric polynomial can thus be associated with the more familiar algebraic polynomial of the same order

$$\mathcal{F}(\zeta) = \sum_{n=0}^{N-1} f_n \zeta^n \tag{3.4}$$

by the transformation

$$\zeta = e^{iu\Omega} \tag{3.5}$$

From the definition of the DFT in (1.47), the coefficients f_n correspond to the samples of a continuous image $f(x)$, which is defined by

$$f(x) = \sum_{n=0}^{N-1} f_n \delta(x - n\varepsilon) \tag{3.6}$$

where ε is the pixel spacing. (3.4) is also known as the Z-transform of an image and can be obtained directly by sampling a continuous image (provided the sampling criteria discussed in §1.5 are met).

Significantly, although a polynomial tends to be thought of as being a function of a real basis variable, it can be transformed to a function of a complex variable simply by allowing the basis variable to be complex. This makes the process of analytic continuation, (Kreysig 1979, p 679), particularly simple.

The reasons for wanting to analytically continue the Fourier transform may not at first be obvious since the process introduces no new information. One advantage of

using analytic continuation is that the transformation given in (3.5) can now be viewed as a conformal mapping between the complex u- and complex ζ-planes, as illustrated in Fig 3.1. (3.5) maps a strip of width $\frac{2\pi}{\Omega}$ in the complex u-plane to the entire complex ζ-plane. As a consequence of the periodic nature of the trigonometric polynomial the consequent algebraic polynomial $\mathcal{F}(\zeta)$ mimics throughout the complex ζ-plane the behaviour of the periodic $F(\dot{u})$ in a strip of width $\frac{2\pi}{\Omega}$ in the complex $\dot{u}$-plane. It is convenient to isolate one period of $F(\dot{u})$, which by convention is defined as the strip from $\dot{u} = -\frac{\pi}{\Omega}$ to $\dot{u} = \frac{\pi}{\Omega}$ called the primary strip, since it is within this strip that $F(\dot{u})$ approximates the continuous Fourier transform $F(u)$. Thus many properties of the DFT can be directly related to those of the Z-transform.

Another advantage of using the Z-transform (or analytically continued DFT) is apparent when equation $\mathcal{F}(\zeta)$ as defined by (3.5), is equated to zero, i.e.

$$\mathcal{F}(\zeta) = \sum_{n=0}^{N-1} f_n \zeta^n = 0 \tag{3.7}$$

Solutions to (3.7) are known as the zeros of $F(\zeta)$. If ζ is restricted to be real, then whilst the maximum number of real zeros is $N-1$, a particular polynomial can have fewer or even no real zeros. When (3.7) is analytically continued into the complex ζ-plane the number of solutions is always exactly equal to $N-1$, although some of the solutions may be repeated. This is the well known fundamental theorem of algebra (Kreysig 1979, p 653). More importantly, a polynomial can be represented in terms of its zeros and the scale factor f_0, i.e.

$$\mathcal{F}(\zeta) = f_0 \prod_{n=0}^{N-1} \left(1 - \frac{\zeta}{\zeta_i}\right) \tag{3.8}$$

Where $\{\zeta_1, \zeta_2, \cdots \zeta_{N-1}\}$ is the set of zeros. These can be plotted in either the complex ζ-plane or complex u-plane in what is called the zero-map of $F(\zeta)$ or $F(u)$ respectively (Sinton et al 1986).

Typically, the zeros of $\mathcal{F}(\zeta)$ lie close to the unit circle, as for example in Fig 3.2a. The distribution of zeros obeys certain symmetries if the coefficients of $\mathcal{F}(\zeta)$ are either real or symmetric. In the former case the zeros are symmetrically distributed about the real axis, so that if $\mathcal{F}(\zeta)$ is zero at ζ_i then it is also zero at ζ_i^*. Note that this does not necessarily imply that zeros on the real axis are multiple. Fig 3.2b illustrates the zeros for a polynomial with real coefficients. When the polynomial is symmetric the zeros are reflected in the unit circle in the complex ζ-plane. Thus if $\mathcal{F}(\zeta_i)$ is zero, $\mathcal{F}(\frac{1}{\zeta_i^*})$ is also zero as indicated in Fig 3.2c.

Using the relationship (3.5) it is possible to convert the zero-maps of $\mathcal{F}(\zeta)$ to the complex $\dot{u}$-plane. Instead of having a finite number of zeros in the entire complex $\dot{u}$-plane the zero map repeats, a result of the periodic nature of $F(\dot{u})$. Fig 3.3 shows the zero maps in the complex $\dot{u}$-plane, corresponding to the ζ-plane zero maps shown in Fig 3.2. The zeros of the continuous $F(u)$ are aperiodic however, a point discussed in detail in §3.4.

3.2 Modelling the visibility using entire functions

When given an arbitrary visibility magnitude it may appear that it is possible to associate any phase with it to reconstruct a feasible object. In practice, however, there is usually

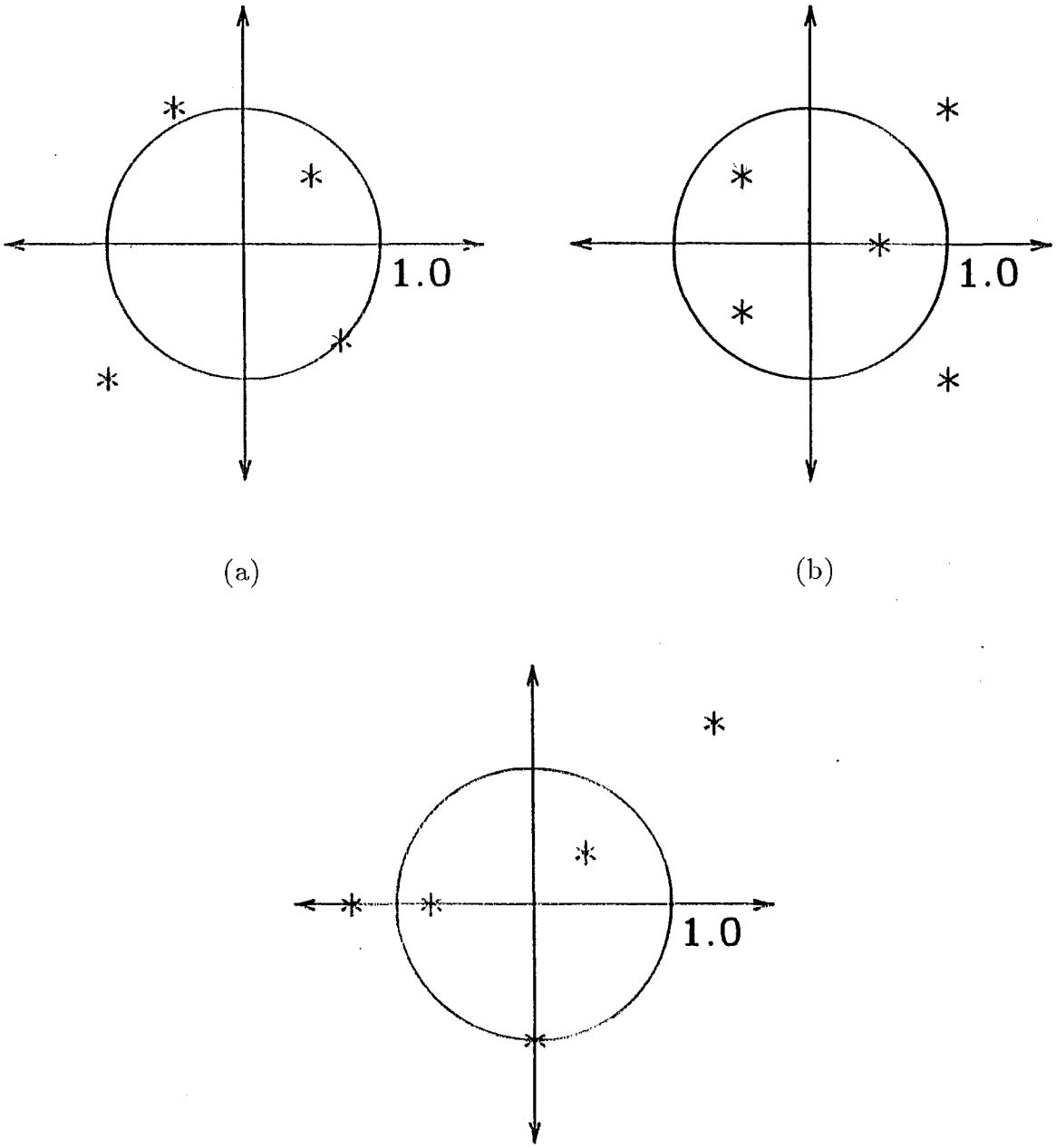


Figure 3.2: Typical zero distributions in Z-space. (a) complex image, (b) real image, (c) symmetric image.

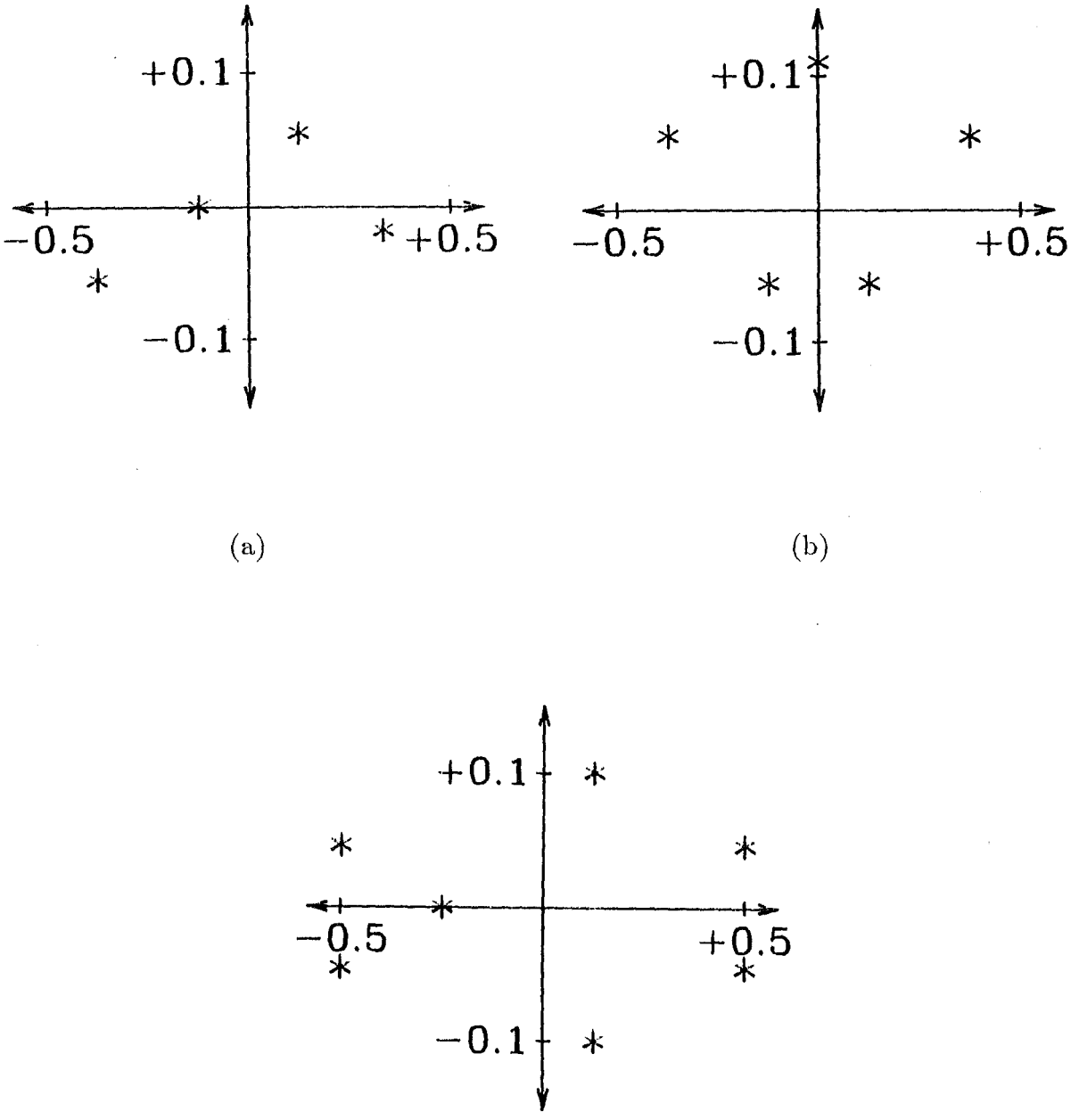


Figure 3.3: Fourier space zero distributions corresponding to the zeros in Z-space shown in Fig 3.2 ($\Omega = 2\pi$). (a) complex image, (b) real image, (c) symmetric image.

some other form of constraint that can be applied, derived from information about the physical nature of the image. The two principal constraints invoked in practice are positivity and a support constraint. The former, positivity, is discussed in §3.6 and applies when the source distribution is known to be both real and non-negative. The latter is of importance because an exact compact support constraint (§1.6) guarantees a relationship between the phase and magnitude of the visibility since the image must vanish outside the region of the support. This can be related to the familiar concept of causality, which demands that no system may respond before an input is applied (Nussenzveig 1972).

Before continuing to discuss entire functions, it is necessary to introduce some more terminology. A "bandlimited" function has traditionally meant a function whose Fourier transform is, in the terminology of this thesis, of exactly compact support (Papoulis 1984). The theory is thus applicable to images which are of finite support, provided the roles of image and Fourier space are reversed.

Another useful concept is the "image-form". The "image-form" differs in image space from the true image by at most a translation and/or an arbitrary phase shift and/or a 180 degree rotation about the coordinate origin. The term "image-form" arises because the appearance (or form) of an image is unaltered by these transformations (Fright and Bates 1982; Bates and McDonnell 1986, §20). Thus if $\vec{x}_1$ and $\vec{x}_2$ are arbitrary constant position vectors and ω_1 and ω_2 are arbitrary real constants then $(f(\vec{x}-\vec{x}_1)e^{i\omega_1})$ and $(f(-\vec{x}-\vec{x}_2)e^{i\omega_2})$ are said to possess the same image form.

Nearly all of the analysis of the one-dimensional phase problem has been directed towards recovering the image-form by expressing the one-dimensional Fourier transform in terms of its Hadamard product. One of the earliest examples of this approach is due to Taylor and Whinnery (1951). The approach taken here follows that of Walther (1963), which still provides one of the simplest expositions of the underlying theory.

Consider a function $f(x)$ which is compact in the interval (a,b). Hence $F(u)$ is by definition an EFET (Requicha 1980) and has an analytic continuation which can be calculated from

$$F(u) = \int_a^b f(x)e^{i2\pi xu}dx \tag{3.9}$$

One can write the Hadamard representation of $F(u)$ as

$$F(u) = e^{(\beta_0+\beta_1 u)}\prod_i^{\infty}\left(1-\frac{u}{u_i}\right)\cdot e^{\frac{u}{u_i}} \tag{3.10}$$

where for simplicity of notation it is assumed that u=0 is not a zero of $F(u)$. Inspection of (3.10) shows that $F^*(u^*)$ is also an EFET. It is convenient to define

$$\begin{aligned} A(u) &= e^{(2\mathcal{R}[\beta_0]+2\mathcal{R}[\beta_1]u)}\prod_i^{\infty}\left(1-\frac{u}{u_i}\right)\cdot\left(1-\frac{u}{u_i^*}\right)\cdot e^{\left(\frac{u}{u_i}+\frac{u}{u_i^*}\right)} && (3.11)\\ &= F(u)F^*(u^*) && (3.12)\\ &= |F(u)|^2 \text{ when u is real} && (3.13)\end{aligned}$$

where $a(x)$ is the autocorrelation of $f(x)$. It is worth emphasising (3.13) which states that $A(u)$ is equal, when u is real, to the square of the magnitude of the Fourier transform of $F(u)$. Since the square of the magnitude of a Fourier transform is real the zeros of $A(u)$ must either be in conjugate pairs or on the real axis. As a consequence of $A(u)$ being positive when u is real it is also necessary for any zeros on the real axis to be of

even multiplicity. If there existed a single zero on the real axis this would imply that the Fourier magnitude was, for some real values of u, negative (Jury 1974). This is, from the definition of the Fourier magnitude, impossible.

It is also worthwhile to consider $a(x)$, the autocorrelation, defined by (1.28), of an arbitrary image $f(x)$. Because $a(x)$ is the Fourier transform of a real function it is by definition symmetric (Papoulis 1984, p 61). Not every symmetric function $s(x)$ is an autocorrelation however, because it is possible for its Fourier transform $S(u)$ to have isolated zeros located on the real u-axis. For example, the rectangular pulse

$$f(x) = \begin{cases} 1 & |x| < \frac{1}{2} \\ \frac{1}{2} & |x| = \frac{1}{2} \\ 0 & |x| > \frac{1}{2} \end{cases} \tag{3.14}$$

can not be an autocorrelation, because its Fourier transform

$$F(u) = \frac{\sin(\pi u)}{\pi u} \tag{3.15}$$

has isolated zeros on the real axis of the complex u-plane at the points

$$u = k\pi \quad k = \pm 1, \pm 2 \ldots \tag{3.16}$$

If, however, an image $f(x)$ and its autocorrelation $a(x)$ are real, then by definition $F(u)$ must also be symmetric about the axis defined by $\mathcal{R}[u] = 0$, which is the imaginary axis in the complex u-plane. Thus the visibility zeros of a real image and its autocorrelation are symmetrically distributed around the real axis.

Denoting by $\overline{F(u)}$ a possible solution of the phase problem one can see that $\overline{F(u)}$ differs from the true visibility $F(u)$, as defined in (3.10), by either the imaginary parts of β_0, β_1 or an arbitrary number of zeros can be replaced by their complex conjugates. Thus

$$\overline{F(u)} = e^{(i\gamma_0 + i\tilde{\gamma}_1 u)} \prod_i^M \left(1 \angle \frac{1 - \frac{u}{u_i}}{1 - \frac{u}{u_i^*}} \right) \cdot F(u) \tag{3.17}$$

where M indicates the number of zeros which have been replaced by their complex conjugates. The term

$$\left(1 \angle \frac{1 - \frac{u}{u_i}}{1 - \frac{u}{u_i^*}} \right) \tag{3.18}$$

is also known as a Blanske factor (Walther 1963). If the function $f(x)$ is known to be real, it is necessary to replace by their conjugates, pairs of zeros reflected in the imaginary axis (i.e u_i and $-u_i^*$ are such a pair of zeros), in order to ensure that the alternative solution $\overline{f(x)}$ is also real where $\overline{f(x)} \longleftrightarrow \overline{F(u)}$.

It should be noted that the factors γ_0 and $\tilde{\gamma}_1$ in (3.17) can not prevent $\overline{f(x)}$ from being compact. The term $e^{i\gamma_0}$ simply scales the function $\overline{f(x)}$ by a complex constant of unit magnitude, whilst $e^{i\tilde{\gamma}_1}$ causes $\overline{f(x)}$ to be translated in image space. Neither of these operations alters the image-form. Similarly replacing every zero by its complex conjugate does not alter the image form, because this is equivalent to reflecting the image in the coordinate origin.

The image-form is however altered if only some of the zeros with non-zero imaginary parts (hereafter called complex zeros) are conjugated. This way of generating

alternative solutions to the phase problem is colloquially known as "zero flipping". If there is an infinite number of complex zeros there is an infinite number of image-forms compatible with a given Fourier magnitude. In practice, however, the constraints of noisy and/or limited data ensure that only a finite number of image-forms can be recovered. Hofstetter (1964) gives several examples of zero-flipping in the absence of noise. Fig 3.4 shows the true image defined by

$$f(x) = \begin{cases} e^{-0.9x} & |x| \leq 1 \\ 0 & |x| > 1 \end{cases} \tag{3.19}$$

The visibility has an infinite number of zeros located at

$$u = \frac{k}{2} - j\frac{0.9}{2\pi} \quad k = \pm 1, \pm 2 \ldots \tag{3.20}$$

as shown in Fig 3.5. Flipping all these zeros results in the mirror image, as shown in Fig 3.4b. Fig 3.4c shows image formed by flipping just the zeros given by $k = 1$ in (3.20), whilst Fig 3.4d shows the effect of flipping zeros for which k is odd in (3.19).

3.3 Relating the Fourier magnitude and phase

The approach described in the previous section has been taken frequently when discussing the ambiguity of the one-dimensional phase problem (cf. Taylor and Whinnery 1951; Walther 1963; Hofstetter 1964; Bates 1969; Burge et al. 1976). Once given an acceptable phase distribution it is then theoretically possible to generate, via zero-flipping, all the possible ambiguities. There remains however the problem of finding an initial feasible solution. Many authors have considered the Hilbert transform relationship (Bates 1969, Nakajima and Asakura 1983a, Ross et al 1978) as a means of relating phase and magnitude. If $a > 0$ in (3.9), then the real and imaginary parts of $F(u)$ are related by the Hilbert transformations:

$$\mathcal{R}[F(u)] = \frac{1}{\pi} Principle \int_{-\infty}^{+\infty} \frac{\mathcal{I}[F(u')]}{u' - u} du' \tag{3.21}$$

$$\mathcal{I}[F(u)] = -\frac{1}{\pi} Principle \int_{-\infty}^{+\infty} \frac{\mathcal{R}[F(u')]}{u' - u} du' \tag{3.22}$$

Because in a practical situation the choice of spatial axes is arbitrary (unlike the choice of the temporal axis), an appropriate choice of origin in image space ensures that $f(x)$ is zero for $x < 0$. As noted by Burge et al. (1976) this ensures that $F(u)$ is analytic in the upper half of the complex u-plane. Hence in practice (3.21) and (3.22) can nearly always be used to relate the real and imaginary parts of the Fourier transform of a spatial function of compact support. Unfortunately the data obtained in practice correspond to the Fourier magnitude and it is not possible to derive the phase directly from (3.21) and (3.22). In order to do so it is necessary to apply the Hilbert transform to a function of $F(u)$ whose real and imaginary parts are independently functions of the magnitude and phase of $F(u)$. Burge et al (1976) have pointed out that the only mathematically tractable function that fulfils this requirement is the logarithm:

$$\ln[F(u)] = \ln[|F(u)|] + i\mathcal{P}(u) \tag{3.23}$$

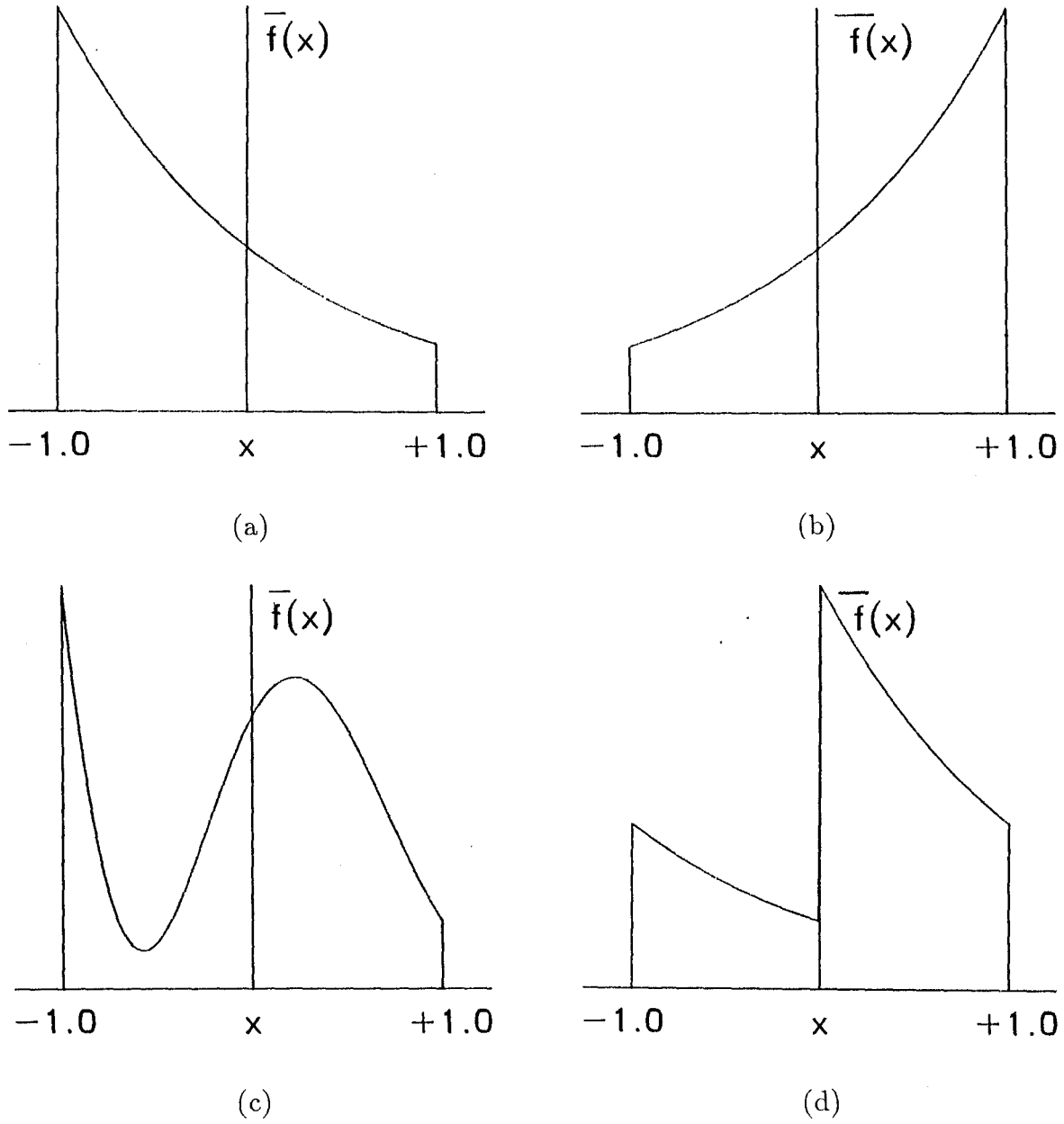


Figure 3.4: Illustration of the effects of zero flipping on a continuous compact image (after Hofstetter (1964)) (a) the true image as defined in (3.19), (b) Image formed by flipping all the visibility zeros in Fig 3.5 (c) image formed by flipping the zeros for which $k = 1$ in (3.20) (d) image formed by flipping the zeros for which k is odd in (3.20)

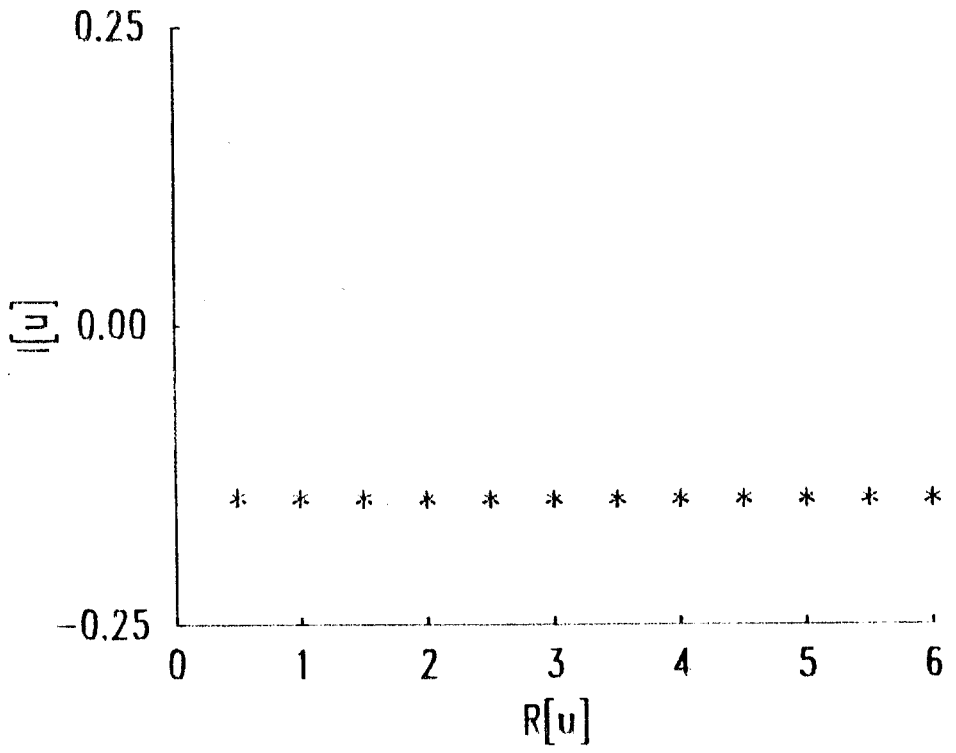


Figure 3.5: Visibility zeros for the continuous image shown in Fig 3.4a. Note only $\mathcal{R}[u] \geq 0$ is shown since $F(u) = F(-u)$

There are a number of difficulties with applying the Hilbert transform to data representing $\ln(|F(u)|)$. Firstly, when $F(u) \to 0$ as $u \to \infty$ (a result of $F(u)$ being of finite energy), the integrals in (3.21) and (3.22) may not converge. Secondly, although $F(u)$ is analytic throughout the complex u-plane, $\ln(F(u))$ is infinite whenever $|F(u)| = 0$. Burge et al (1976) discuss two alternative ways of overcoming this difficulty. Firstly, it may be possible to add a constant A to $F(u)$ so that $\ln(F(u) + A)$ is integrable. In order for this technique to be applicable it is necessary to modify the true image to include a coherent reference (Ross et al 1978), which effectively adds a delta function to the image. If the reference is of sufficiently large magnitude then the visibility can not possess any zeros on the real Fourier axes. In many situations, for example astronomy, it is not possible to modify the image easily.

The second, more feasible, alternative, suggested by Burge et al (1976), does not require that the true image be modified, although it does again require that the lower limit of the support a is greater than zero in (3.9). As a consequence of making $a > 0$, the upper half plane of $F(u)$ is analytic (and contains no zeros). When $a = 0$ the relationship between the magnitude and phase is particularly simple:

$$\ln[|F(u)|] = \frac{u}{\pi} Principle \int_{-\infty}^{+\infty} \frac{\mathcal{P}[F(u')]}{u'(u'-u)} du' + ln[F(0)] \tag{3.24}$$

$$\mathcal{P}[F(u)] = \frac{u}{\pi} Principle \int_{-\infty}^{+\infty} \frac{ln[F(u')]}{u'(u'-u)} du' + \mathcal{P}[F(0)] \tag{3.25}$$

Since the choice of spatial axes is often entirely arbitrary, it is in practice always possible to relate the magnitude and phase of $F(u)$ by (3.24) and (3.25).

There are a number of problems associated with the second of the above approaches. Clearly $\ln(|F(u)|)$ is a rapidly varying function within the vicinity of $|F(u)| = 0$. It is this rapid variation of the integrand which requires one to take its principal value. Bates (1969) has noted that in general it is only possible to accurately compute (3.25) when the allowed variation in $|F(u)|$ is small. Burge et al. (1976) have suggested smoothing the magnitude function before calculating the integral, a process used by Nakajima and Asakura (1982). There are two objections to this approach. Firstly, since detail in the image is reflected in the structure of the Fourier magnitude, any filtering process causes a loss of information in the image. More significantly, for reasons given in chapter 4, filtering of the magnitude can result in there being no image-form compatible with the filtered data.

The susceptibility of the computational evaluation of (3.25) to noise is also of interest. Nakajima (1986) notes that simple minded evaluation of the integral can cause significant errors. It thus appears highly likely that the addition of noise to $|F(u)|$ would significantly alter the estimate of the phase when calculated by (3.25).

Another more subtle objection to using (3.25) to calculate a possible phase distribution stems from the necessity of evaluating the Hilbert integral numerically. In numerical integration it is necessary to approximate a continuous function with a finite number of samples. The inevitably of using a discrete approximation is noted in Napier and Bates (1974) who invoke a model discrete in image space, namely the DFT. This approach has a number of advantages. Firstly, $f(x)$ is likely to be smoother than $\ln(F(u))$, and consequently requires less samples for an adequate representation. Secondly, there is only a finite amount of data in practice, implying that a finite order model is more appropriate. Thirdly, when the data are corrupted by noise, estimating too many model parameters merely models fluctuations that are most likely to be mere random variations in the measurements. The greater robustness of simple models in the presence of noise is widely recognised in the control theory literature (Astrom 1974, Akaike 1978).

Finally, as is noted in §1.6, images of finite energy are effectively compact in both image and Fourier space. Because images compact in both domains can be represented by a finite number of parameters (the coefficients of suitable basis functions) it can be argued that the assumption of a finite order model is, in fact, more appropriate than assuming an exactly compact object in image space.

3.4 Phase retrieval using discrete models

As is noted in §3.1, the DFT is a finite order trigonometric series. Since the series remains the same whether it is represented in terms of the coefficients of a power series or the zeros of a Hadamard product, it is not unreasonable to expect the zeros of the DFT to match those of the continuous Fourier transform in the regions where the DFT is an accurate model of the continuous Fourier transform. Clearly the DFT cannot be expected to be an accurate model of the continuous zeros of the Fourier transform outside the primary strip, but it is worth keeping in mind that if aliasing has occurred (as it must, in practice, to some degree), there may also be significant errors in the location of zeros close to the edges of the primary strip (§3.1, Scivier and Fiddy 1985b).

Another problem that may arise is that the zeros situated at a large distance from the real u-axis may be difficult to locate precisely. If one considers the Fourier transform

of a one-dimensional image analytically continued into the complex ζ-plane, i.e.

$$\int_a^b f(x)e^{i2\pi(\mathcal{R}[u]+i\mathcal{I}[u])x}dx \tag{3.26}$$

the Fourier transform along the line given by $\mathcal{I}[u] = \text{constant}$ is

$$\int_a^b \left(e^{-(2\pi\mathcal{I}[u])x}f(x)\right)e^{i2\pi\mathcal{R}[u]x}dx \tag{3.27}$$

The exponential $e^{-2\pi\mathcal{I}[u]x}$ causes the values of $f(x)$ at the boundaries of the support to contribute more heavily to the calculated transform. Many images are of small amplitude at the edge of the support. By contrast any noise is often more uniformly distributed in image space. Consequently the estimate of the analytically continued Fourier transform is frequently less accurate further from the real u-axis (Sinton 1986). It should be noted that in a few cases analytic continuation can be performed by exponentially filtering the image directly. When the true image is amenable to direct filtering of this type, exponential filtering can be used as the simple basis of a means of phase retrieval (Walker 1981a;b).

Thus, not surprisingly, zeros in the complex plane far removed from the real u-axis may be difficult to locate precisely. Fortunately, by applying the above reasoning in reverse, they do not significantly alter the behaviour of $F(u)$ when u is real. As a result, they do not cause high levels of distortion in the image if they are located inaccurately, provided they are located.

Fig 3.6a shows the sampled approximation to the continuous function defined by (3.19). In this case there is a close correspondence between the zeros of the continuous image, given by (3.20), and those of the DFT approximation, even at quite low sampling rates (Fig 3.7). Increasing the sampling rate increases the width of the primary strip in Fourier space. Whilst it results in more zeros being located at higher spatial frequencies, it does not significantly alter the positions of the zeros found at lower sampling rates.

Assuming the image is sampled as in Fig 3.6a, it is apparent that zero flipping can be applied to the discrete approximation in exactly the same manner as described for the continuous case. Since there are only a finite number of zeros there is, however, only a finite number of image-forms. For a pixellated image comprising N samples, and assuming that all $N-1$ zeros are complex, there are 2^{N-2} possible complex image-forms (since flipping all zeros does not alter the image-form). In the case of real images where the zeros are paired around the imaginary axis there are only $2^{\frac{N}{2}-1}$ possible real image-forms.

If one now flips the pair of zeros closest to the imaginary axis (in Fig 3.7), the image displayed in Fig 3.6b is obtained. Comparison with Fig 3.4c shows a good correspondence between the general shapes of the continuous and the discrete images. Similarly flipping only every other zero in the finite model also produces a good approximation, Fig 3.6c, to that calculated from the continuous model, Fig 3.4d. It is apparent that the discrete models (Figs 3.6b and 3.6c) are inaccurate approximations whenever the continuous images have sharp discontinuities (cf Figs 3.4 and 3.6). The ringing is a natural consequence of modelling a discontinuous function with smooth functions (the complex exponentials). This point is discussed in both §1.5 and Kreysig (1979).

To illustrate the effects of truncation and sampling on a smoother image consider the following Fourier transform pair:

$$f(x) = e^{-\pi(\kappa x)^2} \longleftrightarrow e^{-\pi\left(\frac{u}{\kappa}\right)^2} \tag{3.28}$$

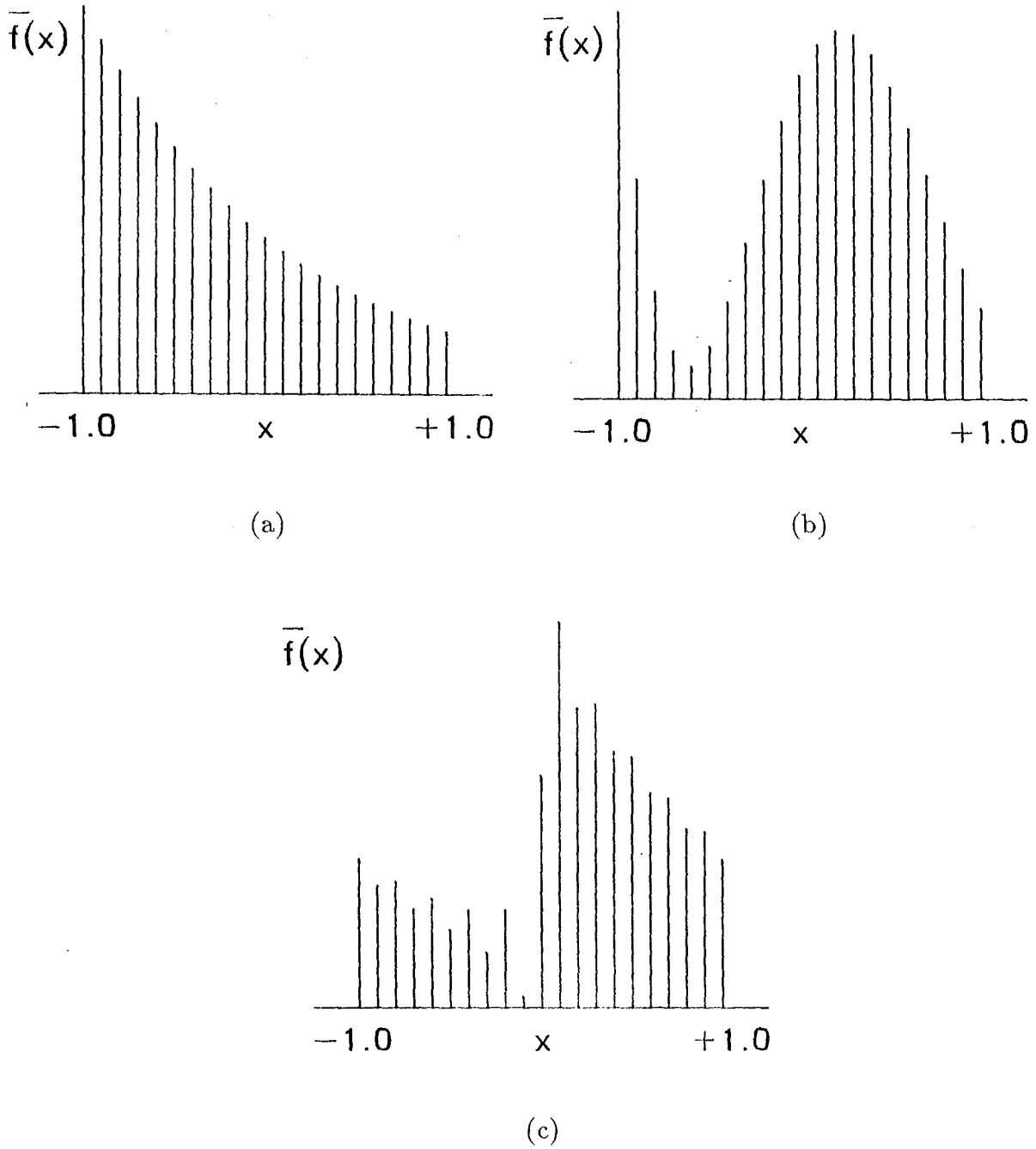


Figure 3.6: Illustration of the effects of zero flipping on a discrete approximation to a continuous image (after Hofstetter (1964)) (a) the sampled approximation of the image in Fig 3.4a (N=21) (b) Sampled image formed by flipping the zeros in Fig 3.7 closest to the imaginary axis, note the good correspondence with Fig 3.4c (c) image formed by flipping every other zero in Fig 3.7. Again note the good correspondence with the continuous image, Fig 3.4d.

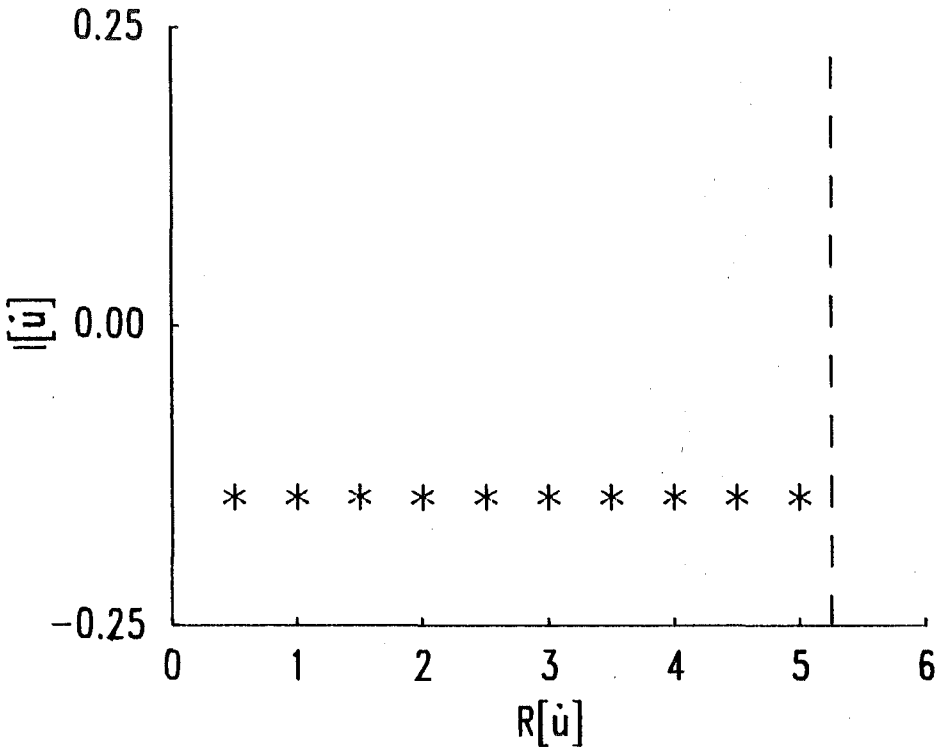


Figure 3.7: Visibility zeros for the discrete image shown in Fig 3.6a. Note only $\mathcal{R}[\dot{u}] \geq 0$ is shown since $F(\dot{u}) = F(-\dot{u})$.

where κ is a positive constant. It should be noted that the continuous $e^{-\pi u^2}$ is never zero in the complex u-plane. Once an image is represented by a finite number of samples, however, it is always possible to compute a finite number of zeros in either the complex ζ-plane or their equivalents in the primary strip of the complex $\dot{u}$-plane. Restricting the number of samples to, for example N, limits the assumed extent of the image to $N\varepsilon = X$, where ε is defined in (3.6). Hence the continuous image modelled is actually of the form

$$f(x) = \begin{cases} e^{-\pi(\kappa x)^2} & |x| \leq X \\ 0 & |x| > X \end{cases} \tag{3.29}$$

where X thus the extent of the (assumed) support in image space. Whereas the Fourier transform of a gaussian is never zero, the visibility of a truncated gaussian does in fact have zeros. This can be seen by separating the truncated exponential into the sum of two functions, as shown in Fig 3.8. The first function is a compact continuous function which is hereafter referred to as the principal. The second function is a rectangular pulse which is in general of much lower energy than the principal.

The Fourier transform of the sum of two functions is given by the sum of their Fourier transforms (Bracewell 1978). Hence at low frequencies the behaviour is dominated by the principal because the total energy of the principal is much larger than that of the pulse, unless X is very small. The behaviour at higher frequencies is more complicated, but can be found from the asymptotic properties of the Fourier transform, which state that if a function is discontinuous in its n^{th} derivative its Fourier transform declines proportionately to $\frac{1}{u^n}$ (Papoulis 1984, p 95). At higher frequencies the Fourier transform of the principal therefore declines proportionately to $\frac{1}{u^2}$ (Papoulis 1984, p 95), because its

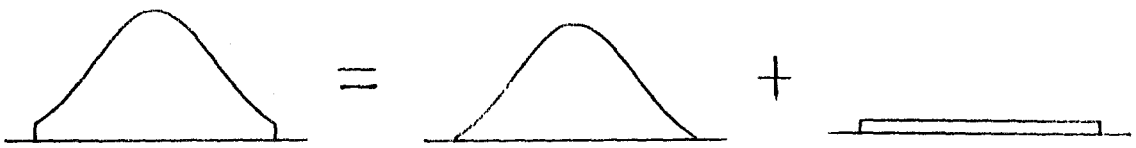


Figure 3.8: Separation of a symmetrically truncated gaussian into a continuous function (referred to in the text as the principal) and a rectangular pulse

first derivative is discontinuous at the edges of the support. By contrast, the rectangular pulse is discontinuous at the edges of the support and thus its visibility declines proportionately to $\frac{1}{u}$. Thus at high frequencies the rectangular pulse is dominant in determining the form of the Fourier transform of the truncated exponential.

Figs 3.9-3.11 shows the effects, on the locations of the zeros of $F(u)$, of changing the sampling rate for different values of κ in (3.29). Because the extent of a visibility in Fourier space is inversely proportional to the image sampling rate in image space, increasing the sampling should reduce the aliasing as well as enabling higher frequencies in Fourier space to be discerned. Because the low frequency zeros remain steadfast when the sampling is increased above a certain level, these zeros in fact correspond to the zeros of the correspondingly truncated continuous gaussian (Ross et al 1978).

3.5 Properties of the one-dimensional phase

In the discussion of homomorphic filtering, presented in §2.6.2, it is noted that this technique relies on forming a continuous single-valued phase function, the so called "unwrapped" phase (Oppenheim and Schafer 1975). In general, however, it is not possible to unambiguously determine the phase of the Fourier transform because it is impossible to distinguish $e^{i2\pi u}$ and $e^{i2\pi u+2k\pi}$ where k is an integer. Whilst this is often overcome by arbitrarily assigning the phase at the centre of Fourier space to zero, it remains impossible to determine a single valued continuous function for the Fourier phase due to ambiguities in the phase function around isolated point zeros.

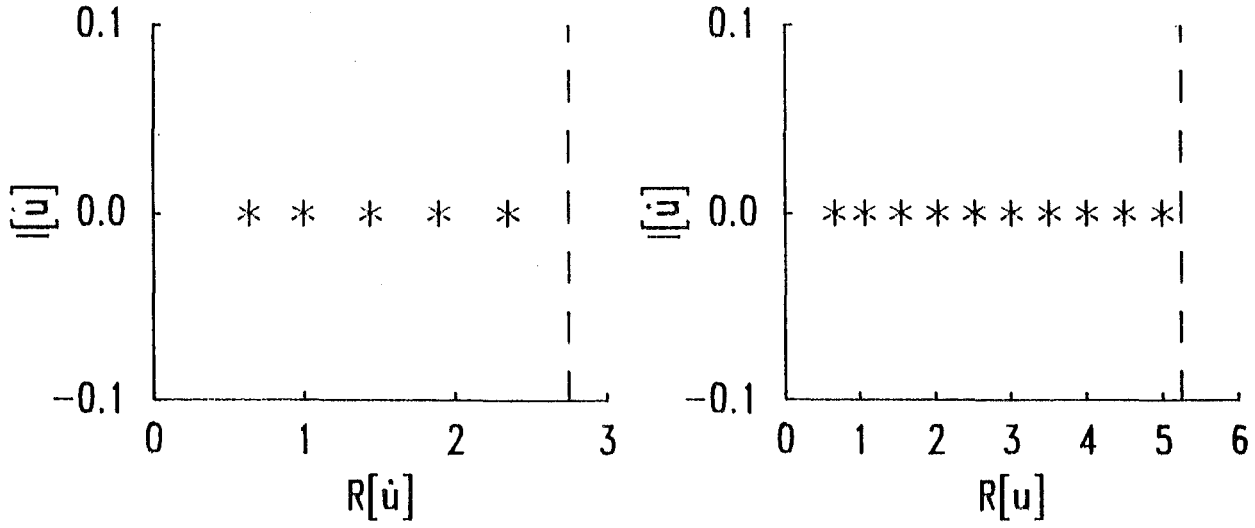
The ambiguity in the phase of the analytically continued Fourier phase can be demonstrated by considering a closed path enclosing an isolated point zero. It is possible to expand $F(u)$ in the vicinity of the zero as

$$F(u + \Delta u) = K\Delta u \tag{3.30}$$

where K is a complex constant. If one now considers a closed path around the point zero defined by

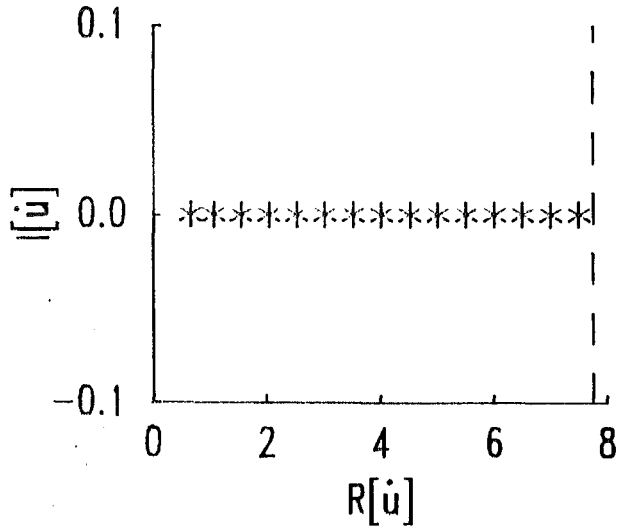
$$\Delta u = e^{i\theta} \quad \theta = -\pi \; to \; \pi \tag{3.31}$$

then if the phase function $\mathcal{P}[F(u)]$ is assumed to be continuous it is apparent that the phase at the starting point increases by 2π for every circuit defined in (3.31). Thus when a path in Fourier space crosses (or encloses) an isolated point zero, it is impossible to



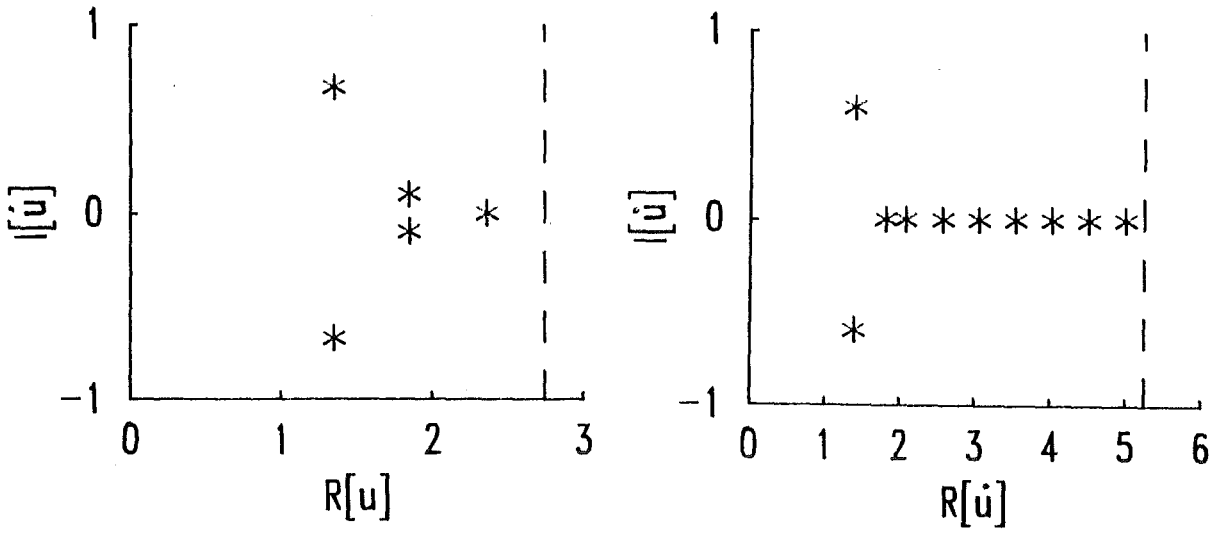
(a)

(b)



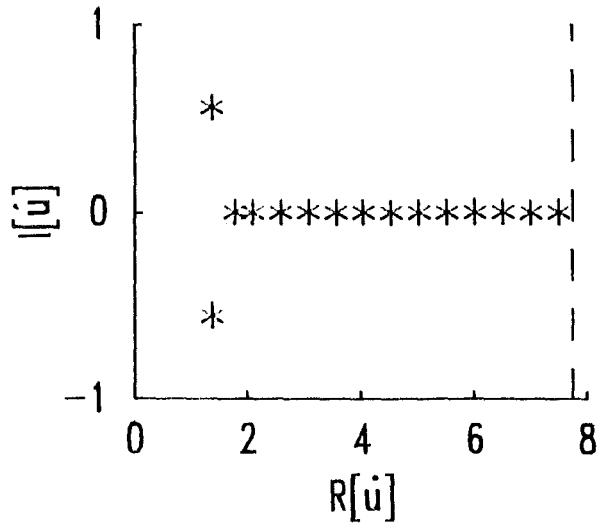
(c)

Figure 3.9: Fourier space zero distributions for the discrete image defined in (3.28) when $\kappa = 0.33$ and X = 1.0. (a) $N = 11$, (b) $N = 21$, (c) $N = 31$.



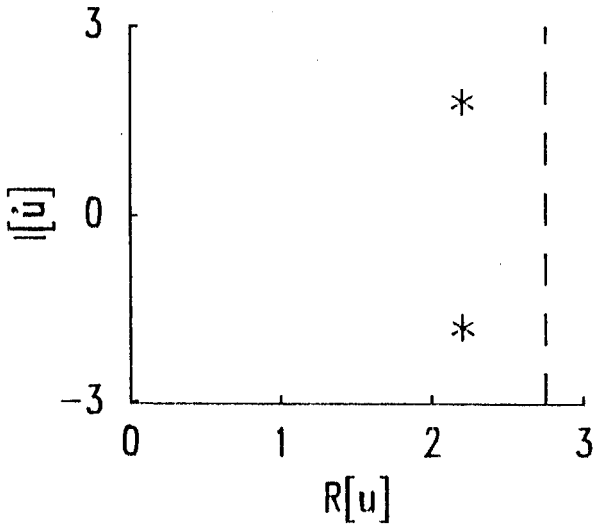
(a)

(b)

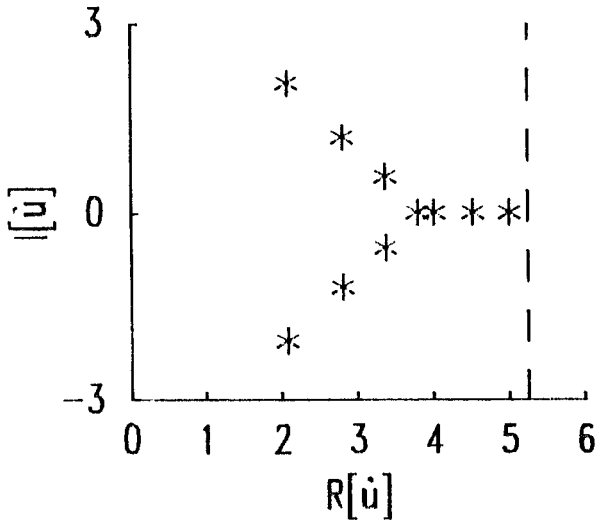


(c)

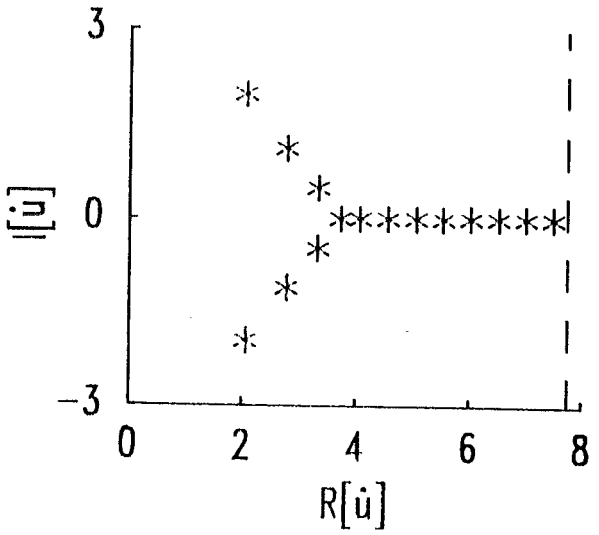
Figure 3.10: Fourier space zero distributions for thediscrete image defined in (3.28) when $\kappa = 0.66$ and X = 1.0. (a) $N = 11$, (b) $N = 21$, (c) $N = 31$.



(a)



(b)



(c)

Figure 3.11: Fourier space zero distributions for the discrete image defined in (3.28) when $\kappa = 1.0$ and X = 1.0. (a) $N = 11$, (b) $N = 21$, (c) $N = 31$.

define a single-valued continuous phase function (Oppenheim and Schafer 1975, Tribolet 1977). In one-dimension it can happen that along an open path between two different points in Fourier space there are no isolated zeros, hence it is often possible to "unwrap" the Fourier phase along a line in Fourier space, for example, the real u-axis.

The theoretical solution to the dilemma of defining a continuous phase function over the complex u-plane is to represent the phase function as an infinite Riemann surface (Kreysig 1979), with branch cuts terminating at the zeros. A practical alternative, and the technique employed in this thesis, is to define $\mathcal{P}[F(u)]$ to exist only within the interval $[-\pi, \pi]$ i.e. modulo 2π. The existence of discontinuities in the phase function does not affect the phase difference between arbitrary points in Fourier space, provided that this phase difference is also defined appropriately (a point discussed further in §6.2).

3.6 Using Image Positivity as a Constraint

§§3.3 and 3.4 establish that the relationship between a visibility's magnitude and phase is strongly constrained when the image is compact. If the image is also known to be positive, further constraints can be imposed on the relationship.

It is of interest to determine how the positivity constraint imposed in image space can be related to an equivalent set of constraints on the visibility. This problem has been dealt with extensively in crystallography where it is at the heart of the "Direct Methods" pioneered in the 1950's by Karle (1986) and Hauptman (1986). As noted in §2.4, in crystallography the visibility is a discrete function with its sampling rate determined by the periodic structure of the crystal. This section adopts the notation of Karle (1986) where an arbitrary sample in (K-dimensional) Fourier space is denoted by F_{k1} with the sample at the centre of Fourier space being denoted by F_0.

A necessary and sufficient condition for a periodic image, $p(\vec{x})$, to be positive is that

$$\begin{vmatrix} F_0 & F_{-k_1} & F_{-k_2} & \cdots & F_{-N} \\ F_{k_1} & F_0 & F_{k_1-k_2} & \cdots & F_{k_1-N} \\ F_{k_2} & F_{k_2-k_1} & F_0 & \cdots & F_{k_2-N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_N & F_{N-k_1} & F_{N-k_2} & \cdots & F_0 \end{vmatrix} \geq 0 \tag{3.32}$$

where F_{k_n} are located at the Nyquist samples in Fourier space (Karle 1986). It is important to realise that, since the visibility of a real image is conjugate symmetric, the above determinant is always real. It is possible to derive a number of inequalities from (3.32), for example the third order inequality,

$$\begin{vmatrix} F_0 & F_{-k_1} & F_{-k_2} \\ F_{k_1} & F_0 & F_{k_1-k_2} \\ F_{k_2} & F_{k_2-k_1} & F_0 \end{vmatrix} \geq 0 \tag{3.33}$$

which can be rewritten as

$$\left| F_{k_2} - \frac{F_{k_1} F_{k_2-k}}{F_0} \right| \leq \frac{\left| \begin{matrix} F_0 & F_{-k_1} \\ F_{k_1} & F_0 \end{matrix} \right|^{\frac{1}{2}} \cdot \left| \begin{matrix} F_0 & F_{k_1-k_2} \\ F_{k_2-k_1} & F_0 \end{matrix} \right|^{\frac{1}{2}}}{F_0} \tag{3.34}$$

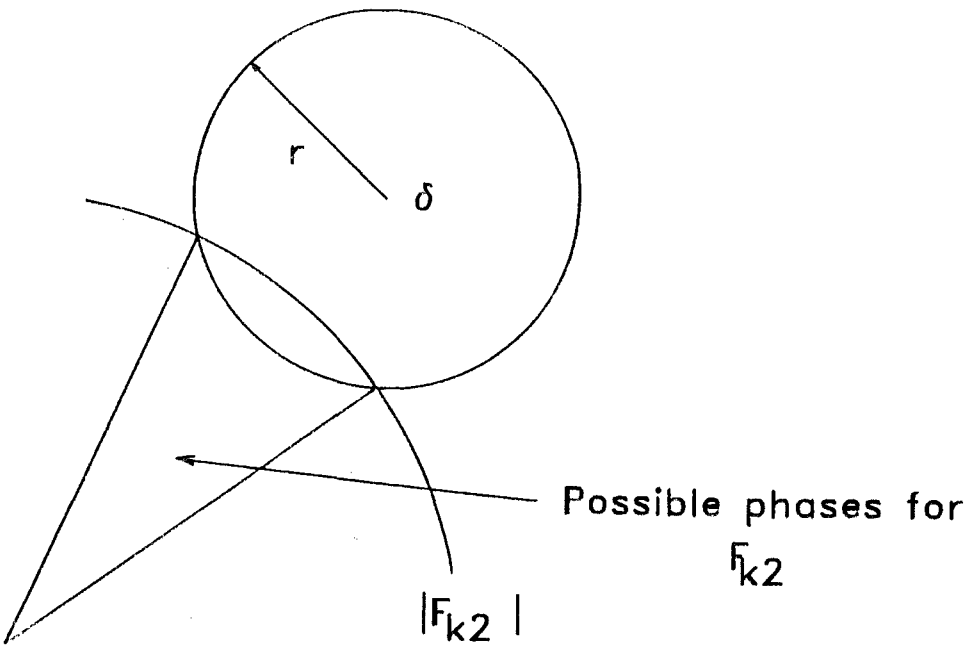


Figure 3.12: Demonstration of how $\mathcal{P}[F_{k_2}]$ is constrained to lie within a certain range by (3.36) and the knowledge of $|F_{k_2}|$.

When F_0, F_{k_1} and F_{k_2} are all large in (3.34) the right hand side, denoted hereafter as r, becomes small. As a consequence

$$F_{k_2} \approx \frac{F_{k_1} F_{k_2-k}}{F_0} = \delta \tag{3.35}$$

a formula which can be used to relate the phases of the Nyquist samples of the Fourier transform.

Alternatively one can rewrite (3.35) in the form

$$|F_{k_2} - \delta| \leq r \tag{3.36}$$

Fig 3.12 shows how (3.36) and the knowledge of $|F_{k_1}|$ constrains the phase of F_{k_2}. In order to develop a working algorithm, it is useful to think of (3.36) in terms of a probability density function (Karle 1986).

There are other methods of using image positivity as a constraint on the Fourier phase, for example Cocke (1985) used the Cauchy-Schwarz inequality whilst Cadzow and Sun (1986) have noted that it is possible to use the following variation of Parseval's theorem.

$$\int_0^{2\pi} f(x)|w(x)|^2 dx = \sum_{m,n=-\infty}^{\infty} F(m-k)W(m)W^*(k) \geq 0 \tag{3.37}$$

Positivity can also be used to reduce the number of possible solutions found by zero flipping, because many of these images may contain negative parts and can be rejected as possible solutions (Napier 1971, Napier and Bates 1974).

3.7 Deconvolution using zero based representations

One of the major advantages of a representation in terms of zeros is that it serves as a starting point for a means of deconvolution. Consider $\mathcal{G}(\zeta)$ the Z-transform of the image $g(x)$, which is the convolution of two images $f(x)$ and $h(x)$. By the convolution theorem $\mathcal{G}(\zeta) = \mathcal{F}(\zeta)\mathcal{H}(\zeta)$. If one now considers a point ζ' in the complex ζ-plane, it is not possible to infer from $\mathcal{G}(\zeta')$ precise information about the values of either $\mathcal{F}(\zeta')$ or $\mathcal{H}(\zeta')$ except when $G(\zeta') = 0$. This is because when $G(\zeta') = 0$, then either $\mathcal{F}(\zeta')$ or $\mathcal{H}(\zeta') = 0$. Since it is possible to reconstruct either $\mathcal{F}(\zeta)$ or $\mathcal{H}(\zeta)$ from their respective zeros the problem of deconvolving $G(\zeta)$ is reduced to one of partitioning the zeros of $G(\zeta)$ into those resulting from $\mathcal{F}(\zeta)$ and $\mathcal{H}(\zeta)$ respectively.

Polynomials have finite numbers of zeros and thus there is only a finite, albeit possibly very large, number of ways of partitioning the zeros of $G(\zeta)$. Specifically, for a sampled image of length $2N + 1$ samples, there are $2N$ zeros and consequently 2^{2N} possible functions for $f(x)$ and $h(x)$ (which may be reduced if zeros are repeated). When dealing with EFETs factorisation is less well defined. An example of the problems that can arise in factoring entire functions is discussed by Sanz and Huang (1985) and Lawton and Morrison (1987). Consider the Z-transform of the image $f(x)$ which is assumed to be representable by two samples, each of unit magnitude (i.e. the pixellated image represented by {1,1}).

$$G(\zeta) = 1 + \zeta \tag{3.38}$$

Since $G(\zeta)$ has only one zero it is obviously impossible to partition it between $\mathcal{F}(\zeta)$ and $\mathcal{H}(\zeta)$. Hence $G(\zeta)$ is irreducible, i.e. it can not be factored into polynomials of lower order. If one now considers the trigonometric polynomial associated with the image $g(x)$ as an entire function

$$G(u) = 1 + e^{ju} \tag{3.39}$$

clearly this has an infinite number of zeros occurring at $u = 2n\pi + \pi$ $(n = 0, \pm1, \pm2, \ldots)$. Since there are an infinite number of zeros it can be factored in an infinite number of ways into other entire functions, for example

$$G(u) = (1 + je^{j\frac{u}{2}})(1 - je^{j\frac{u}{2}}) \tag{3.40}$$

The zeros of $(1 + je^{j\frac{u}{2}})$ occur at $4n\pi + \pi$ $(n = 0, \pm1, \pm2, \ldots)$ and those of $(1 - je^{j\frac{u}{2}})$ at $4n\pi - \pi$ $(n = 0, \pm1, \pm2, \ldots)$ The union of the zeros of these two factors are the zeros of $(1 + e^{ju})$ and thus $G(u)$ is irreducible.

It is instructive to consider this difference in a practical context. The function $(1 + je^{\frac{ju}{2}})$ corresponds to a pixellated image {1, j} sampled at twice the rate of the original convolution. If one assumes that the original sequence was sampled adequately with respect to the Nyquist criterion, the factors have an extent in Fourier space which is larger than that of the convolution and thus violate the assumption that the initial sampling of the convolution was adequate. Limiting the model to polynomials, rather than using more general entire functions, helps prevent these spurious solutions to a deconvolution problem.

Chapter 4

TWO-DIMENSIONAL MODELLING

The previous chapter discusses how the zeros of a one dimensional polynomial can be used to relate the phase and the magnitude of the visibility of an image of compact support. In two dimensions the relationship between Fourier magnitude and phase would at first sight appear more complicated. In this chapter, however, it is shown that there is almost always a unique relationship between the multi-dimensional Fourier magnitude and phase (in contrast with the ambiguous relationship in one-dimension described in §3.2). The question of uniqueness is perhaps the most important theme of this chapter.

This chapter considers only two-dimensional images, although the procedures used can easily be generalised to higher dimensional images. The techniques employed involve reducing a single two-dimensional Fourier transform pair to a set of related one-dimensional Fourier transform pairs, by a process known as projection. There exists a one-dimensional Fourier transform between a projection in image space and a line in the two-dimensional Fourier transform plane. This one-dimensional transform can be treated by the techniques presented in chapter 3.

§4.1 discusses "angular projections" which Napier and Bates (1974) propose as a means of of simplifying the two-dimensional Fourier transform. The form of projection they suggest is however incomplete, because it is not possible to reconstruct a unique two-dimensional Fourier transform from the set of one-dimensional Fourier transforms obtained by angular projections. §4.1 concludes with a generalisation of the angular projection procedure which can be used to completely characterise the two-dimensional Fourier transform.

For reasons given in §3.4 a discrete model of the Fourier transform is again employed. By using the Z-transform it is then possible to employ a simpler form of projection than the angular projection given in §4.1, and it is this projection which is employed for the remainder of the chapter. §4.2 discusses projections and their intimate relation to the concept of phase closure in Fourier space. As a conclusion to §4.2, a method of phase retrieval, based on phase closure is described. Although using phase closure to effect phase retrieval was initially proposed by Bates (1982), it was first implemented by Deighton et al (1985). Although the method is, in its present form, computationally impractical it provides an excellent introduction to the differences between one- and two-dimensional Fourier transforms.

§4.3 addresses the question of whether there is a unique relationship between the magnitude and phase of the visibility of a compact image. The approach taken follows that

of Bruck and Sodin (1979) and Hayes (1982) where the Fourier spectrum is approximated by a two-dimensional polynomial rather than by an exact entire function model. When using a polynomial model it is particularly simple to show why there is often only one image-form compatible with a given Fourier magnitude.

As with the one-dimensional phase problem (chapter 3) a large number of researchers have chosen to employ EFET's to model the visibility of an exactly compact object in more than one dimension (Kiedron 1981; Lawton 1981; Manolitsakis 1982; Sanz and Huang 1983b; Stefanescu 1985; Sanz 1985b; Lawton and Morrison 1987). The theoretical analyses of these authors provides a good understanding of the properties of the two-dimensional phase problem and is discussed in §4.4.

In practice, however, one is faced with a limited amount of corrupted data. In the author's opinion this fact alone precludes the use of models which require, by definition, an infinite number of parameters. §4.5 thus returns to the polynomial model of the Fourier spectrum introduced in §4.2. An important concept introduced in this section is that of the zero-sheet (Lane et al. 1987; Lane and Bates 1987a), which generalises the uniqueness arguments of Napier and Bates discussed in §4.1.

The zero-sheet is the complete two-dimensional extension of the zero-map of a one-dimension visibility. It is shown that the point zeros of the one-dimensional projections of a two-dimensional Z-transform usually form part of a single continuous, analytic surface which is called in this thesis a zero-sheet. The zero-sheet of an analytically continued N-dimensional Fourier transform can be modelled by a surface of $2(N-1)$ dimensions, existing in a $2N$ dimensional space. Thus the zero-sheet of a two-dimensional image is characterised by a two-dimensional surface existing in a four-dimensional space. Only when this zero-sheet is the union of several distinct zero-sheets is it possible to express the corresponding image as a convolution of smaller compact images. Display of these zero-sheets presents considerable problems (associated with their visualisation) which are described in §4.6. §4.6 also introduces the algorithmic basis used for mapping the zero-sheets.

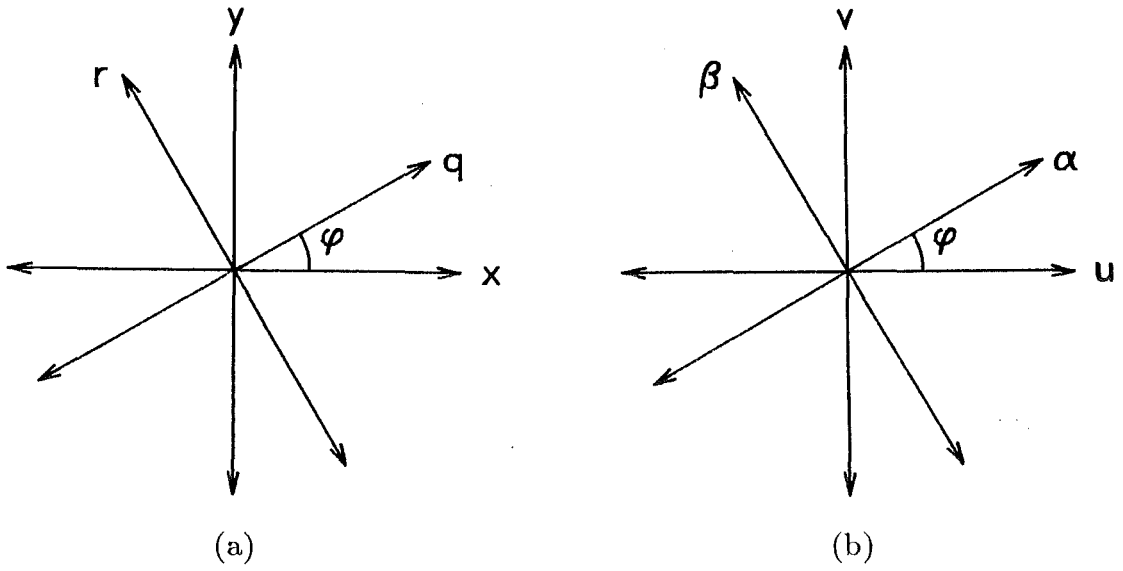
Although the zero-sheets were initially formulated by a generalisation of the technique of Napier and Bates (1974), it is possible to view them as a special case of the zero manifolds of EFETs. The theory presented in §4.4 provides confirmation of the practical results obtained by mapping the zero-sheets of two-dimensional polynomials. §4.7 discusses this and how zero-sheets can be used to effect both phase retrieval and blind deconvolution of quite large images.

Recovering an image from its zero-sheet is the subject of §4.8. Although it is in theory always possible to recover an image from its zero-sheet in practice there are a number of difficulties. §4.8 introduces two main techniques for recovering an image from its zero-sheet. The first relies on the application of phase closure to relate the Fourier transform derived along orthogonal lines in Fourier space. The second, due to Curtis et al (1985), formulates the image recovery problem in terms of a system of linear equations. §4.8 describes both techniques.

The final section of this chapter, §4.9, addresses the important problem of how noise affects zero-sheets. It is shown that noise effectively causes the multiple zero-sheets of a convolution to become linked. This provides a geometric complement to the algebraic arguments of other authors (Hayes 1982; Sanz et al. 1983).

4.1 Reduction of two-dimensional Fourier transforms to one-dimensional Fourier transforms

One of the first practical attempts to reduce the two-dimensional Fourier phase problem to a set of related one-dimensional problems was the use of the projection theorem (Napier and Bates 1974). The projection theorem, which forms the basis of computed tomography (Garden 1984), relates a set of line integrals calculated in image space to the Fourier transform along a line in Fourier space.



(a) (b)

Figure 4.1: The coordinate syatems used for describing angular projections and their Fourier transforms. (a) image space (b) Fourier space.

The theorem arises quite naturally from considering co-ordinates in image and Fourier space rotated by an angle φ (Fig 4.1). If one now considers the Fourier transform relationship in the rotated coordinates:

$$F_\varphi(\alpha,\beta) = \iint f_\varphi(q,r)e^{i2\pi(\alpha q+\beta r)}dr\,dq \tag{4.1}$$

where the coordinates q and r are defined such that

$$f_\varphi(q,r) = f(x\cos\varphi + y\sin\varphi, y\cos\varphi - x\sin\varphi) \tag{4.2}$$

where the subscript φ is used to indicate that the (q,r) and (α,β) axes are rotated by an angle φ to the (x,y) and (u,v) coordinates respectively. In computed tomography an angular projection is defined by

$$s_\varphi(q) = \int f_\varphi(q,r)dr \tag{4.3}$$

which corresponds to a set of line integrals indicated in Fig 4.2. If β is set to zero in (4.1) the two-dimensional Fourier transform reduces to

$$F_\varphi(\alpha,0) = \int f_\varphi(q,r)dr e^{i2\pi\alpha q}dq \tag{4.4}$$

$$= \int s_\varphi(q)e^{i2\pi\alpha q}dq \tag{4.5}$$

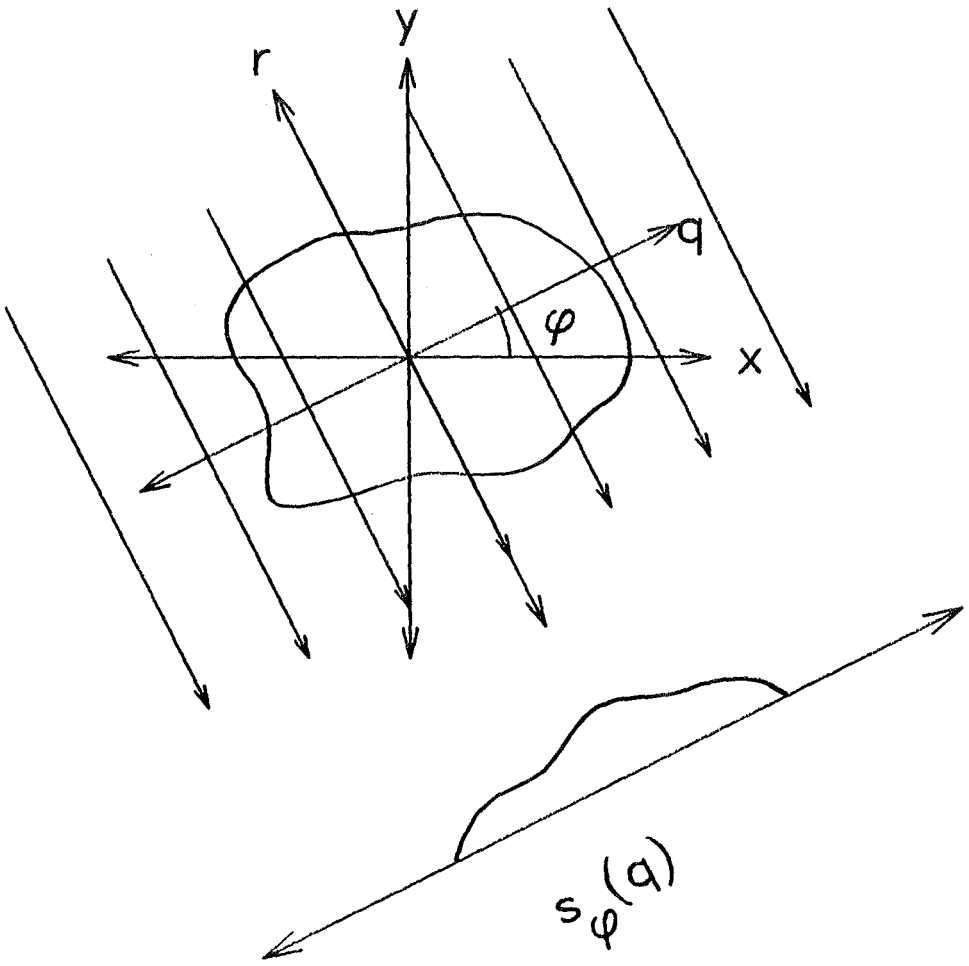


Figure 4.2: The formation of an angular projection by a series of parallel line integrals

There thus exists a one-dimensional relationship between the angular projection in image space and the Fourier transform along a line passing through the coordinate origin in Fourier space. The one-dimensional polynomial representation of this transformation enables the Fourier phase and magnitude to be related by the techniques discussed in chapter 3. If there was no relation between the infinite number of one-dimensional projections which are formed by varying φ, then relating the two-dimensional Fourier magnitude and phase would be a hopeless task, as each individual projection would be individually subject to the ambiguities present in the one-dimensional case.

Fortunately, because $F(u,v)$ is an entire function, the zeros of the angular projection's visibility $F_\varphi(\alpha,0)$ must vary continuously with φ (Napier and Bates 1974). It is then possible to relate the point zeros of $F_\varphi(\alpha,0)$ to those of $F_{\varphi+\Delta\varphi}(\alpha,0)$, where $\Delta\varphi$ is a small quantity. Further application of continuity dictates that the point zeros of $F_\varphi(\alpha,0)$ must follow continuous paths in the analytically continued α-plane. Points along these paths are identified by the parameter φ (as shown in Fig 4.3). Thus when the starting zero-map is selected by the zero flipping process (§§3.2–3.4) the zero paths formed by varying φ can be used to determine the zero-maps corresponding to other values of φ. It is apparent from this reasoning that the Fourier phase problem is at least no more ambiguous in two-dimensions than in one.

The ambiguity can be further reduced if the image is known to be positive (Napier and Bates 1974), since this means that all one-dimensional projections must also be positive. Because positivity is enforced over a set of linked one-dimensional projections, rather than a single one-dimensional image, positivity can be expected to be a more

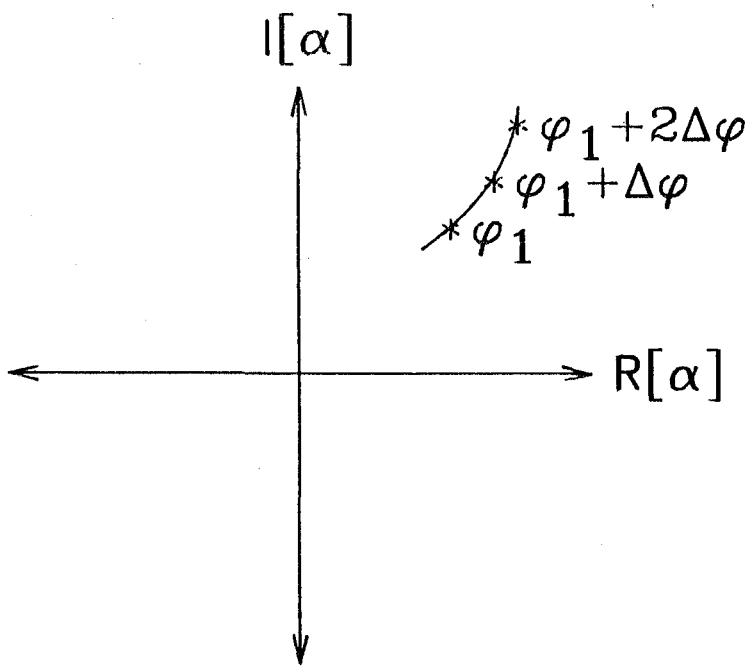


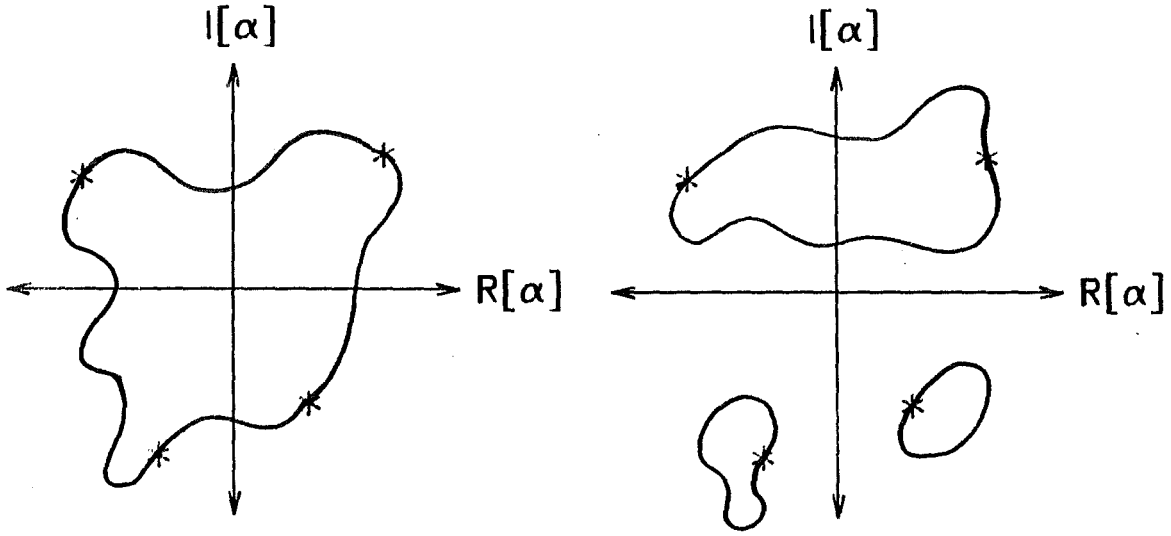
Figure 4.3: Migration of a point zero of $F_\varphi(\alpha, 0)$ as φ is increased incrementally

powerful constraint in two-dimensions (Napier and Bates 1974).

Since the zero maps of $F_\varphi(\alpha, 0)$ can be related directly to $\mathcal{P}[F_\varphi(\alpha, 0)]$, it is possible to generate a two-dimensional phase distribution for $F(u, v)$. This phase distribution is continuous, although allowance must be made for isolated zeros in the (u, v)-plane (a point discussed in §3.5). The point $(0, 0)$ is common to all the visibilities of all the one-dimensional projections and thus can be used to align the phases of the visibilities of the one-dimensional projections at all values of φ.

When the procedure outlined in the previous paragraph is used to generate a two-dimensional phase distribution, in general only one choice of starting zero map (and of course its mirror image in the real axis) yield a compact image-form. One possible reason for this behaviour is that, because the projections formed at φ=0 and $\varphi = \pi$ are equal, the zero maps of the projection's visibilities should be the same. By again appealing to the entire nature of $F(u, v)$ the paths followed by the individual zeros of $F_\varphi(\alpha, 0)$ must then form closed paths (hereafter called zero-contours) when φ is varied from 0 to π, Fig 4.4a. There are two possible ways to form these closed paths: either a zero follows a path leading to its original position or it forms a path terminating at another of the initial zero positions. When the point zeros of the initial projection are linked by zero contours it is apparent that they must be treated as an inseparable collection if $F(u, v)$ is entire. If the zeros of a zero-map are linked by zero-contours they must be treated as a group.

Fig 4.4 shows a hypothetical zero-map for an image. Superimposed on this figure (Figs 4a and 4b)
are some idealised zero contours formed by varying φ from 0 to π. Since the point zeros of this initial zero-map are linked into a single inseparable collection by the zero contours, it is apparent that they cannot be "flipped" independently. As a consequence there are no other possible image-forms that can be created by the process of zero-flipping.



(a) (b)

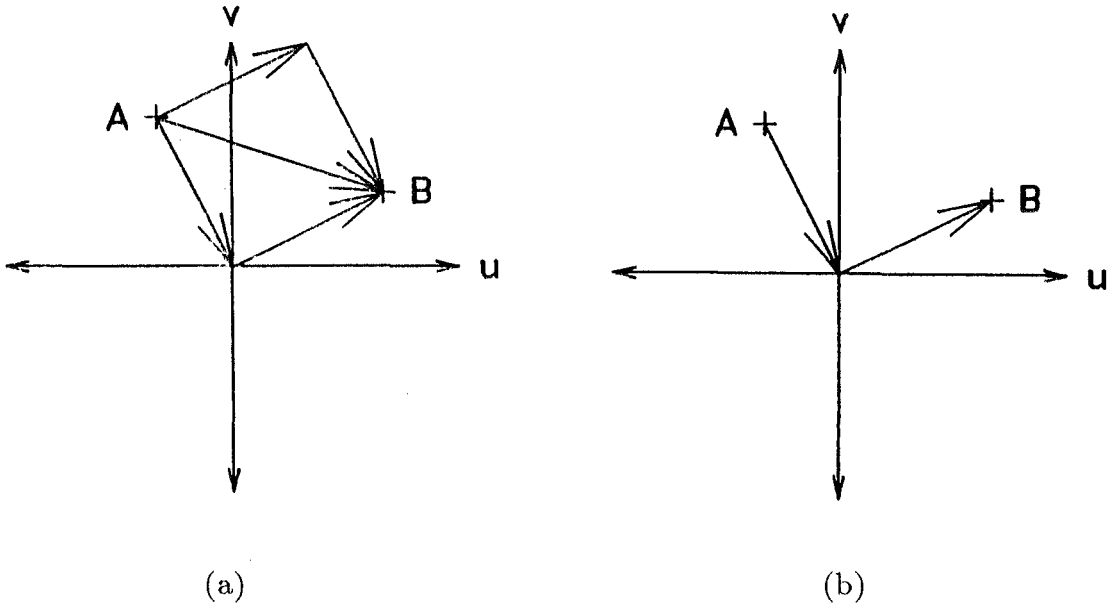
Figure 4.4: Closed zero-contours formed by varying φ from 0 to π. $*$'s are used to indicate the zeros of $F_0(\alpha, 0)$. (a) all zeros of $F_0(\alpha, 0)$ linked by a single zero-contour (b) partial linkage of the zeros of $F_0(\alpha, 0)$ by multiple zero-contours

Initial numerical experiments, performed by the author, showed that this zero linkage did in fact occur, but although the zeros often did form inseparable collections there was usually more than one. Even when a zero collection is "flipped" together the resulting image-form is not necessarily compact. It is therefore apparent that the set of angular visibilities formed by varying φ from 0 to π does not completely characterise the two-dimensional visibility from which they were derived.

In §3.1 the one-dimensional Fourier transform is analytically continued into the complex plane. The analytically continued Fourier transform exists in a space spanned by two real dimensions and enables a visibility to be represented by its possibly complex zeros. Using just the real zeros of the one-dimensional visibility does not provide a complete representation of the one-dimensional Fourier transform.

Applying similar reasoning to the two-dimensional Fourier transform results in an analytically continued transform existing in a four-dimensional space. It is apparent the set of angular projections defined by (4.3) exist in a space which is a function of only three variables $\mathcal{R}[\alpha]$, $\mathcal{I}[\alpha]$ and φ. The zeros of the set of angular projections thus can not provide a complete representation of the analytically continued two-dimensional Fourier transform, a function which resides in a four-dimensional space.

The difference between one-dimensional and two-dimensional Fourier transforms is also apparent when considering the phase difference between two arbitrary points A and B. In a two-dimensional Fourier transform it is possible to calculate the phase difference between two arbitrary points A and B by applying continuity along an infinite number of paths Fig 4.5a. By contrast in a two-dimensional Fourier transform built up from a set of angular projections there is only one way of calculating the phase difference between points A and B, Fig 4.5b. Since the visibilities of all angular projections pass through the origin of Fourier space it is not possible to form closed circuits.



(a) (b)

Figure 4.5: Possible paths for calculating the phase difference between two points A and B in Fourier space. In the general two-dimensional Fourier transform shown in (a), there are an infinite number of possible paths. By contrast when the Fourier transform is built up from angular projections (b), there is only one possible path.

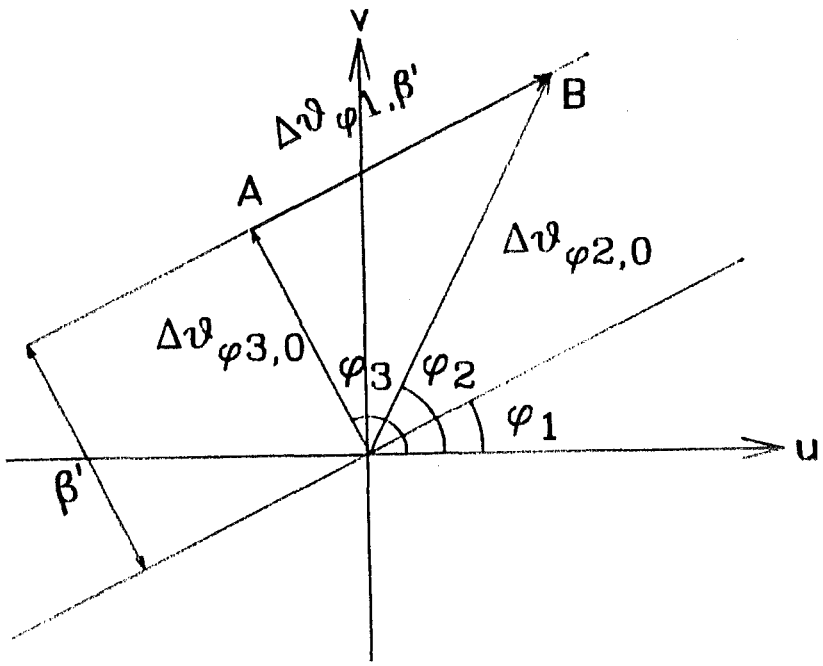


Figure 4.6: Phase closure using generalised angular projections. Note the phase difference (modulo 2π) between any two points in Fourier space must be the same.

It is, however, possible to use a generalised form of projection

$$s_{\varphi,\beta}(q) = \int f_{\varphi}(q,r)e^{i2\pi\beta r}dq \tag{4.6}$$

a form hereafter referred to as the general angular projection. Since

$$F_{\varphi}(\alpha,\beta) = \int s_{\varphi,\beta}(q,r)dr\,e^{i2\pi\alpha q}dr \tag{4.7}$$

there exists a one-dimensional Fourier transform relationship between the general angular projection and a line displaced from the origin in Fourier space in the manner shown in Fig 4.6. Since the visibilities of general angular projections do not pass through a single point in Fourier space it is possible to form closed paths in Fourier space. If $F(u,v)$ is entire the phase difference between points A and B must agree, modulo 2π, independently of the path by which it is calculated. Thus

$$\Delta\vartheta_{\varphi_1,\beta'} \ mod\ 2\pi = (\Delta\vartheta_{\varphi_2,0} + \Delta\vartheta_{\varphi_3,0}) \ mod\ 2\pi \tag{4.8}$$

The 2π ambiguity arises because of the possibilty of a singularity in the (u,v)-plane a point which is discussed in detail in §3.5.

4.2 Projections and phase closure

The previous section notes that the calculated difference between arbitrary points in Fourier space must be independent of the path taken. This requires the phase difference around a closed path to always equal 0 (modulo 2π), a requirement known as "phase closure" (Fright 1984). For reasons given in §3.4 it is convenient to employ a discrete model of an image. The image is thus represented by an M x M array of pixels (some of which may be zero), e.g.

$$f(x,y) \approx \sum_{m,n=0}^{M-1} f_{m,n}\delta(x - m\alpha)\delta(y - n\alpha) \tag{4.9}$$

where α is the pixel spacing in both the x and y directions. Using $\overset{z}{\longleftrightarrow}$ to indicate a Z-transform pair where the sampling conditions described in §1.5 are met,

$$f(x,y) \overset{z}{\longleftrightarrow} \mathcal{F}(\zeta,\gamma) = \sum_{m,n=0}^{M-1} f_{m,n}\zeta^m\gamma^n \tag{4.10}$$

It should be noted that the order of a two-dimensional polynomial is given by the highest sum of the indices of both basis variables present. Hence the order of the Z-transform in (4.10) is equal to $(2M-2)$.

The convolution theorem also applies in Z-space where it is written in two-dimensions as

$$\mathcal{G}(\zeta,\gamma) = \mathcal{F}(\zeta,\gamma)\mathcal{H}(\zeta,\gamma) \tag{4.11}$$

$$= \sum_{i=0}^{2M-2}\sum_{j=0}^{2M-2}\left(\sum_{m=0}^{M-1}\sum_{n=0}^{M-1} f_{m,n}h_{i-m,j-m}\right)\zeta^i\gamma^j \tag{4.12}$$

The process of convolution in image space is thus equivalent to polynomial multiplication in Z-space. In order to relate this two dimensional polynomial to the periodic approximation of the continuous Fourier transform it is necessary to employ the transformations (§3.1),

$$\zeta = e^{(i2\pi u\Omega)}, \qquad \gamma = e^{(i2\pi v\Omega)} \tag{4.13}$$

where Ω, the fundamental frequency in Fourier space (Requicha 1980), can be found directly from α and M using (1.47) and (1.49).

From examination of (4.10) it is a simple matter to reduce $F(\zeta,\gamma)$ to a one-dimensional Z-transform, all that is required is to fix the (complex) value of either ζ or γ. If for example, ζ is fixed at a constant value denoted by ζ_0, then the resulting one-dimensional projection is given by

$$\mathcal{F}_{\zeta_0}(\gamma) = \sum_{m=0}^{M-1} \left(\sum_{n=0}^{M-1} f_{m,n}\zeta_0^m \right) \gamma^n \tag{4.14}$$

which is a one-dimensional polynomial in γ.

In order to reconstruct a two-dimensional image using the inverse Fourier transform it is necessary to have the values of $F(u,v)$ when u and v are both real. From (4.13) u and v are only real when ζ and γ are both of unit magnitude. Hence in order to find the Fourier transform along a line parallel to the v-axis given by $u = u_0$, it is first necessary to fix ζ at $e^{i2\pi u_0}$. The value of $F(u_0,v)$ is then given by the Z-spectrum on the unit circle in the complex γ-plane.

In practice the values of the Fourier transform obtained along lines defined by $u = u_0$ are only known to within a complex constant of unit magnitude (due to the limitations of zero-flipping discussed in §§3.2 and 3.4). As a consequence it is not possible to deduce phase variation perpendicular to the v-axis using solely ζ-projections. This phase variation can be obtained, however, by interchanging the roles of γ and ζ in the preceding discussion. This gives the phase variation along lines parallel to the u-axis which is, by definition, orthogonal to the v-axis.

It is thus possible to apply phase closure along closed rectangular paths (referred to as circuits). These closed circuits comprise one-dimensional projections formed by fixing both ζ and γ at complex constants of unit magnitude. If one now considers the circuit and notation defined in Fig 4.7 it is apparent that

$$(\Delta\theta_{\zeta_0} + \Delta\theta_{\gamma_0}) \ mod \ 2\pi = (\Delta\theta_{\zeta_1} + \Delta\theta_{\gamma_1}) \ mod \ 2\pi \tag{4.15}$$

Bates (1982b) suggests that it should be possible to generate all possible values of the above phase differences by one-dimensional zero-flipping (§3.4). The algorithm then selects those phase values which do in fact satisfy (4.15). Repeating the process for other closed circuits enables the phase distribution for all of Fourier space to be calculated. Unfortunately, for a general complex image the number of phases that has to be tested grows exponentially with the size of the image. For a general complex M x M image there are 2^{2M-3} possible phases for each of $\Delta\theta_{\zeta_0} + {}_{\Delta}\theta_{\gamma_0}$ and $\Delta\theta_{\zeta_1} + \Delta\theta_{\gamma_1}$. Clearly the number of comparisons that are required increases dramatically with the image size. Napier and Bates rejected the algorithm as being beyond the computational power available at the time. Deighton et al (1985) produced an algorithm which recovered a 16 x 16 real image but did not describe their algorithm. During the course of investigating this method of phase

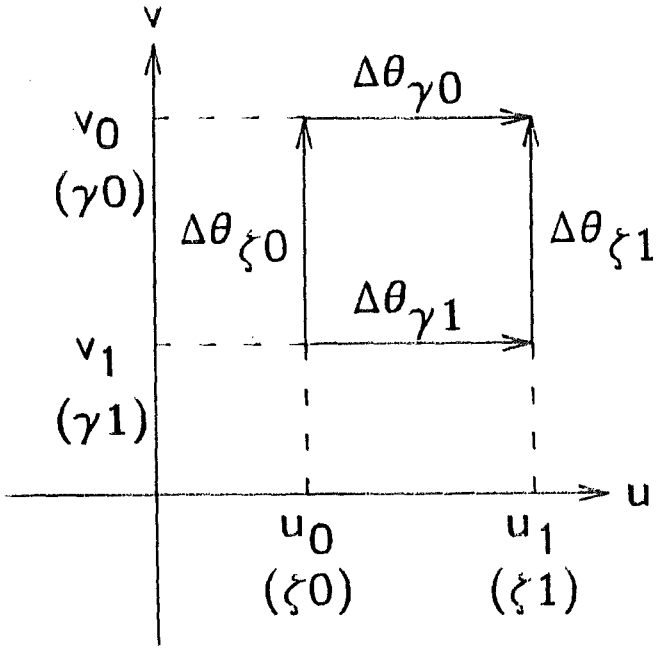


Figure 4.7: Phase closure in Fourier space using projections in Z-space. Note the magnitudes of ζ_0,ζ_1,γ_0 and γ_1 are all equal to 1.

retrieval, the author developed an optimised algorithm which is capable of recovering a 16 x 16 real image, within 15 minutes on a VAX 11-750 minicomputer.

The above algorithm relies on generating and storing all possible values of $\Delta\theta_{\zeta_0} + \Delta\theta_{\gamma_0}$, by using zero flipping on the one-dimensional projections. In practice to distinguish between an image and its reflection in the coordinate origin one complex zero in the ζ_0 projection is fixed. This has the additional advantage of reducing the number of possible values for the phase difference in Fourier space. An indication of which zeros are flipped to produce the phase difference is also stored. A similar list, except that all zeros are flipped, is made of $\Delta\theta_{\zeta_1} + \Delta\theta_{\gamma_1}$.

Searching for identical values in the two lists, by comparing each element from one with all the values in the other list would clearly be a hopeless task. In order to find the matching pair in an acceptable time the two lists are sorted in order of increasing phase and merged into one list. The sorting was achieved efficiently using Quicksort (Horowitz and Sahni 1978), a commercial implementation of which is available on the VAX 11-750 computer which was used. Since the list is in ascending order of possible phases it is now only necessary to compare adjacent values in the list of phase differences. If there is only one pair of phase differences that correspond to different paths in Fourier space then it possible to determine the phase along all projections which comprise the circuit by using the associated information about which zeros have been flipped to generate the phase variation along a projection.

The mapping of Fourier space is then completed by imposing phase closure on different rectangular circuits in Fourier space. Provided two sides of these circuits are along the original projections there is a vastly reduced number of possible phase differences, because zero flipping need be only applied to the new projections. Because the

time taken to perform phase closure varies exponentially with the number of zeros to be flipped, it takes far less time to perform phase closure around new circuits.

In practice there may remain a small number of possible phase differences after calculation of phase closure on the initial rectangular circuit. This final ambiguity can also be resolved by applying phase closure around a different circuit in Fourier space. Provided the new circuit is bounded by some of the original projections used in the original circuit, the number of possible phase differences for the new circuit is again much reduced and the remaining ambiguity can be resolved in much less time than is required to compute phase closure on the original circuit.

The algorithm is unfortunately unworkable in its present form when noise is added to the Fourier modulus. This noise causes a shift in positions of the zeros of the visibility. Hence the exact zeros of the true image are no longer an exact subset of the zeros of the autocorrelation. As a result it is no longer possible to calculate the phase differences in (4.15) exactly. Hence it is not possible to rely upon exact phase closure in the presence of noise. There is a strong possibility of improving the algorithm, along the line indicated in chapter 7.

4.3 Uniqueness of two-dimensional phase retrieval

The previous section shows that phase closure can be used as the algorithmic basis for phase recovery. It does not, however, rigorously show that there is a unique relationship between the magnitude and phase of the Fourier transform. The question of uniqueness is addressed in pioneering work by Bruck and Sodin (1979) and Hayes (1982), who indicate that there should in fact be a unique relationship between the magnitude and phase of the visbility of a compact discrete object.

It is well known that if the coefficients $g_{m,n}$ of a two-dimensional polynomial are chosen arbitrarily then it is impossible to factorise the given polynomial into the product of lower order polynomials (Huang et al. 1971). This is often referred to as the lack of a fundamental theorem of algebra in more than one-dimension (Coolidge 1959). Only when $g(x,y)$ is a convolution is it possible to write $\mathcal{G}(\zeta,\gamma)$ as the product of two polynomials, as in (4.11).

Bruck and Sodin (1979) were the first to realise that the absence of a fundamental theorem of algebra in more than one-dimension had significance for the Fourier phase problem. As an autocorrelation is a particular form of convolution (1.25) then it must be reducible (i.e. factorisable). For example the Z-spectrum of $ff(x,y)$ can be written

$$\mathcal{FF}(\zeta,\gamma) = \mathcal{F}(\zeta,\gamma)\mathcal{F}^*\left(\frac{1}{\zeta^*},\frac{1}{\gamma^*}\right) \tag{4.16}$$

where $\mathcal{F}^*(\frac{1}{\zeta^*},\frac{1}{\gamma^*})$ is the Z-transform of $f^*(-x,-y)$ (Oppenheim and Schafer 1975). Hence the Z-spectrum of an autocorrelation has at least two factors, one corresponding to the true image and the other to the image conjugated and rotated in the coordinate origin.

Hayes and McClellan (1982) formalised this argument by rigorously proving that the reducible polynomials formed a set of measure zero in the set of multidimensional polynomials of a given order. The approach taken was to assign the coefficients of the multidimensional polynomial to a vector (cf Nowinski 1981). It can be shown that the dimension of the space spanned by reducible polynomials of a given order is less than the dimension of the space spanned by irreducible polynomials of the same order.

	ζ^0	ζ^1	ζ^2
γ^0		1	1
γ^1	1	$3+\delta$	1
γ^2	1	1	

$$\mathcal{F}(\zeta,\gamma) = 1 + \zeta + \gamma + (3+\delta)\zeta\gamma + \zeta\gamma^2 + \zeta^2\gamma + \zeta^2\gamma^2$$

Figure 4.8: Simple pixellated image. Note that the Z-spectrum is only factorisable when $\delta = 0$.

It is also possible to apply the absence of a fundamental theorem of algebra in more than one-dimension to blind deconvolution (Lane and Bates 1987a), with phase retrieval considered as a special case. When an image $g(x,y)$ can be expressed as the convolution of two other images (or components) then the Z-spectrum must be factorisable. The reverse process, that of deconvolution, is thus equivalent to factorising the Z-spectrum. Referring to the definition of the two-dimensional Z-transform in (4.10), it is apparent that factorisation of the Z-spectrum is equivalent to factorising the two-dimensional polynomial used to represent $\mathcal{G}(\zeta,\gamma)$ (Lane and Bates 1987a).

Since polynomials form a ring of factorisation (Sanz and Huang 1985), the polynomial factors of $\mathcal{G}(\zeta,\gamma)$ can be always used to reconstruct discrete compact images. These images are sampled at the same value of α (refer (4.10)) as the original convolution. Assuming the sampling of the convolution is adequate, if it is not possible to factorise $\mathcal{G}(\zeta,\gamma)$ then it is not possible to form discrete compact objects which can be convolved together to produce the observed $g(x,y)$.

The above approach is simplified when viewed from a "degrees of freedom" approach, a method taken by Sanz and Huang (1985). Consider the convolution of an N x N pixel image $f(x,y)$ with an M x M pixel image $h(x,y)$. From (4.12) the resultant convolution is thus $(N+M-1)$ x $(N+M-1)$ pixels. It is possible to associate a nonlinear equation of the pixels of $f(x,y)$ and $h(x,y)$ with each pixel of the convolution. When N and M are both greater than 1 the total number of unknowns (i.e. the number of pixels in $f(x,y)$ and $h(x,y)$) is always less than the number of equations (i.e. the number of pixels in the convolution).

When dealing with an arbitrary $(N+M-1)$ x $(N+M-1)$ pixel image there is in general no solution to this overdetermined set of equations. As a consequence it is not possible to find two non-trivial component images which can be convolved together to equal this arbitrary choice of image.

The dimensionality approach can be taken further to show that the property of irreducibility is stable (Hayes 1982). Hence when the coefficients of an irreducible polynomial are perturbed the resultant polynomial is again almost always irreducible. By contrast, perturbing the coefficients of a reducible polynomial almost always results in an irreducible polynomial (Sanz 1985a). As an example of this consider the pixellated image shown in Fig 4.8. Simple algebra shows that this image is only factorisable when δ is equal to zero. The effects of perturbation of the polynomial coefficients on the reducibility of a polynomial is an important issue because all practical measurements are contaminated by noise. Thus even when the sampled data model a convolution, the two-dimensional polynomial representation is in general irreducible. Deconvolution (or phase

retrieval) thus involves finding the reducible polynomial "closest" to (i.e the one which best approximates) the irreducible polynomial.

The application of arguments concerning reducibility to the phase problem is a somewhat cloudy issue. Although statistically unlikely, a reducible polynomial models the processes of both convolution and autocorrelation. Hence it is quite feasible for the true visibility to be reducible if in fact it is formed by a convolutional process.

The effect on the phase problem of the true image being a convolution is shown in Figs 4.9 and 4.10. Fig 4.9a shows the true image $g(x,y)$ which is the convolution of the individual images $f(x,y)$ and $h(x,y)$ shown in Figs 4.9b and 4.9c respectively. It is readily apparent that

$$\begin{aligned} |G(u,v)| &= |F(u,v)||H(u,v)| && (4.17)\\ &= |F^*(u,v)||H(u,v)| && (4.18)\\ &= |F(u,v)||H^*(u,v)| && (4.19)\\ &= |F^*(u,v)||H^*(u,v)| && (4.20) \end{aligned}$$

which correspond to the images shown in Fig 4.10. Since Fig 4.10a is equivalent to 4.10b rotated 180 degrees around the coordinate origin they both relate to the same image-form. By similar reasoning 4.10c and 4.10d also have the same image-form, but this image-form is clearly different to the image-form shown in Figs 4.10a and 4.10b. In general, if the true image is the convolution of N components then there are 2^N image-forms (Lane et al 1987).

4.4 Using entire functions of exponential type

Multidimensional EFETs exhibit a number of the properties of multi-dimensional polynomials although there are some significant differences. One major complication is that a general EFET can only be described by an infinite number of parameters. Hence it is not possible to show that there are more irreducible than reducible EFETs. In fact it can be argued that the reverse is in fact true (Manolitsakis 1982).

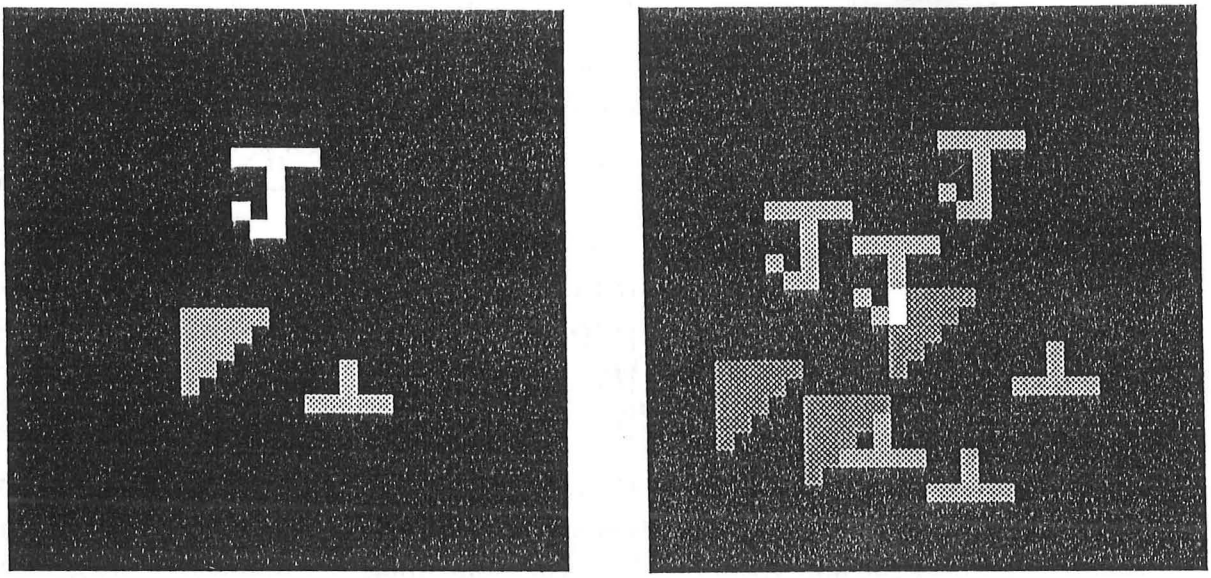
If an EFET is assumed irreducible then it is possible to show that there is a unique relationship between its magnitude and phase (Sanz et al. 1983). The conditions for irreducibility of an EFET are however different from those of a polynomial as is discussed in §3.7. It is possible, however, to analyse the points where the EFET is zero,

$$F(u,v) = 0 \qquad (4.21)$$

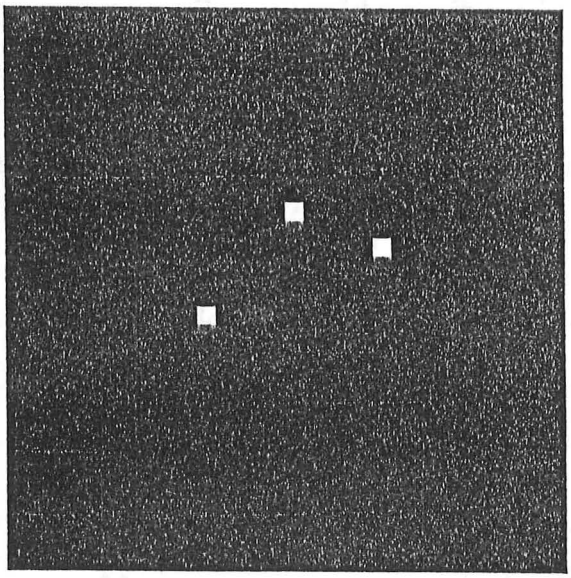
The use of $F(u,v)$ is appropriate as EFETs are exact models of the visibility of a continuous compact object (unlike polynomials which merely approximate it over a finite period). As noted by Manolitsakis (1982) the set of points where $F(u,v)$ is zero, hereafter denoted by Υ, forms a continuous analytic set. In other words the set of points Υ forms a connected Riemann surface, provided $F(u,v)$ is not constant (Lawton and Morrison 1987). If $F(u,v)$ is irreducible then it is not possible to form analytic subsets Υ_1 and Υ_2 such that

$$\Upsilon = \Upsilon_1 \cup \Upsilon_2 \qquad (4.22)$$

(Manolitsakis 1982; Sanz and Huang 1983b, Lawton and Morrison 1987). These analytic subsets each correspond to factors of the EFET and hence to components of a convolution

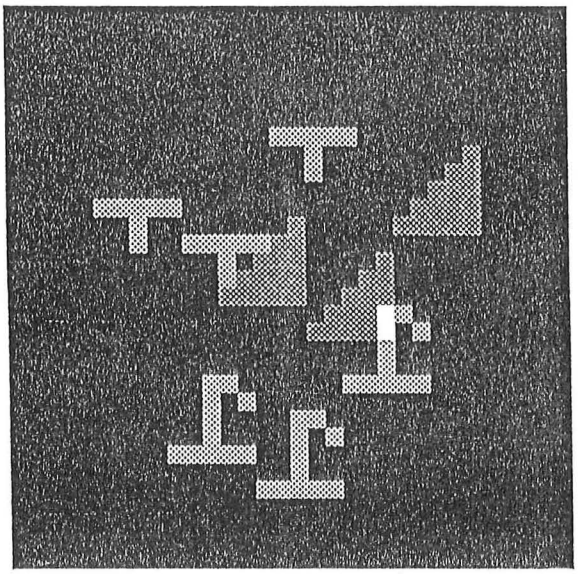


(a) (b)

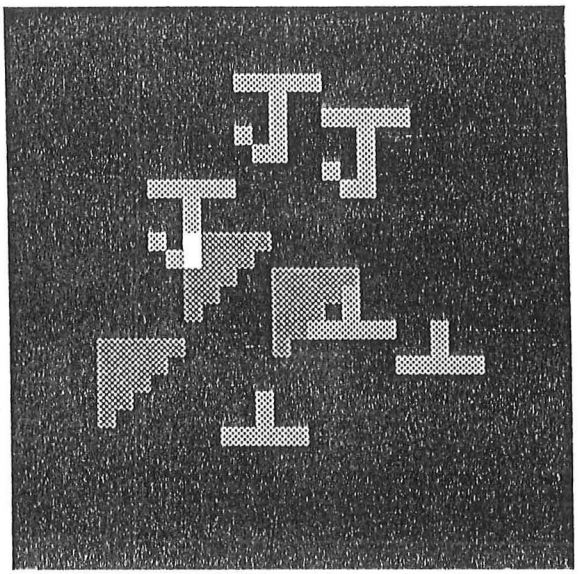


(c)

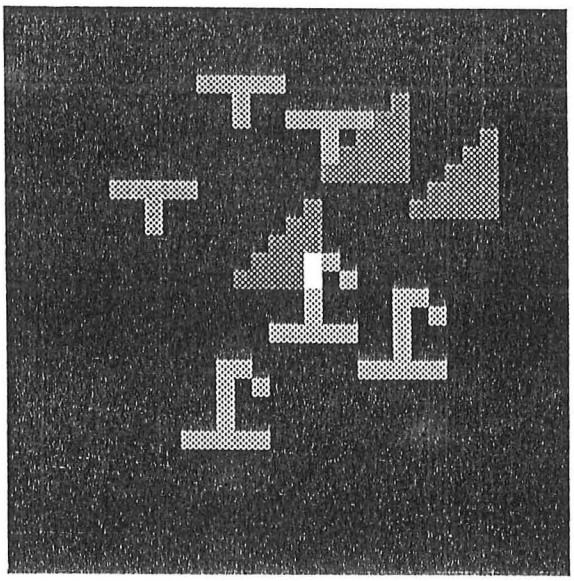
Figure 4.9: The convolution of two real simple objects. (a) $g(x,y) = f(x,y) \odot h(x,y)$ (b) $f(x,y)$ (c) $h(x,y)$.



(a)



(b)



(c)

Figure 4.10: Alternative image-forms whose visibility is the same as that of Fig 4.9a
(a) $f(-x,-y) \odot h(x,y) \longleftrightarrow F^*(u,v)H(u,v)$
(b) $f(x,y) \odot h(-x,-y) \longleftrightarrow F(u,v)H^*(u,v)$
(c) $f(-x,-y) \odot h(-x,-y) \longleftrightarrow F^*(u,v)H^*(u,v)$.

in image space. It should be noted that these differ from the factors of a polynomial in one crucial aspect. There is no guarantee that the factors are themselves EFETs since the set of entire functions does not form a ring of factorisation. Thus, unlike polynomials, they may not correspond to physically reasonable solutions. Furthermore it is not always possible to generate other possible image-forms using the simple process of zero-flipping (Stefanescu 1985).

The question of irreducibility of two-dimensional entire functions is to a large degree still unanswered, although it is known that two-dimensional trigonometric polynomials are irreducible when considered to be entire functions (Sanz and Huang 1983b). This is in contrast to the equivalent situation in one-dimension discussed in §3.7.

There also exist some results on the irreducibility of functions generated by rotating symmetric functions around their point of symmetry (Lawton 1981). These functions are essentially one-dimensional and can be solved using a modification of the one-dimensional theory.

4.5 Zero-sheets of two-dimensional polynomials

It is also possible to analyse the analytic sets where a multidimensional polynomial $\mathcal{F}(\zeta,\gamma)$ is zero, a set of points denoted in this thesis by $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$. The recovery of a two-dimensional polynomial from its zero-sheet is theoretically possible (Curtis and Oppenheim 1987, Curtis et al. 1985) but presents certain practical difficulties (Sanz 1985b) which are discussed further in §4.8. For the present, however, it is assumed that it is possible to generate $\mathcal{F}(\zeta,\gamma)$ from $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$.

The compactness of $f(x,y)$ ensures that $\mathcal{F}(\zeta,\gamma)$ is analytic for all finite values of the complex variables ζ and γ. In order to describe $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$ it is convenient to rewrite (4.10) as

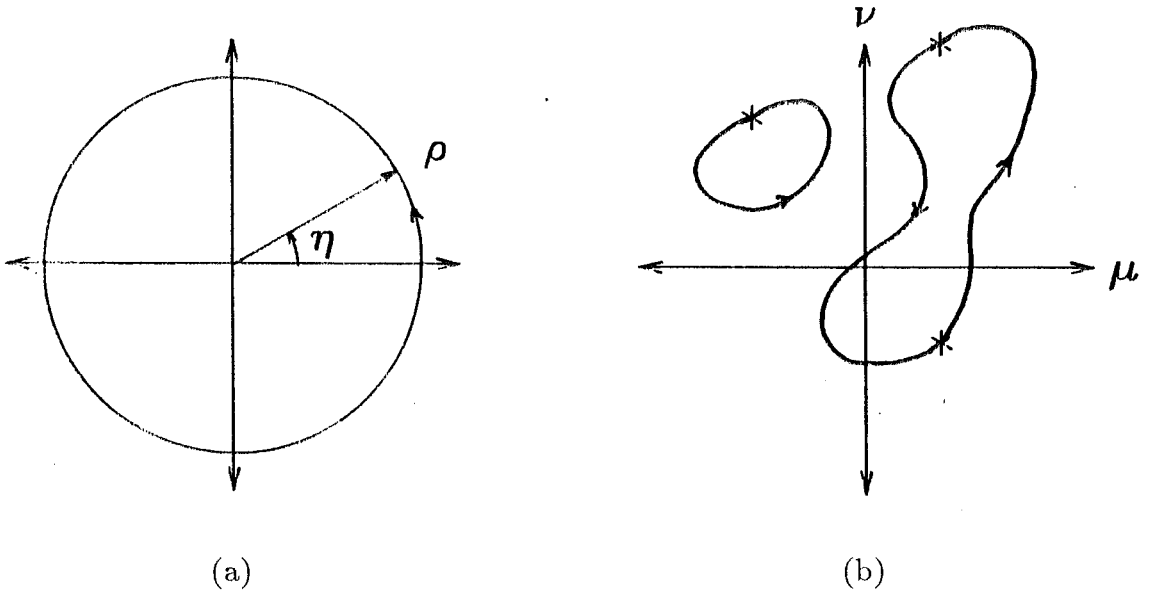
$$\mathcal{F}(\zeta,\gamma) = A(\zeta)e^{i\psi(\zeta)} \prod_{\ell=1}^{M} (\gamma - \overline{\gamma_\ell}(\zeta)) \tag{4.23}$$

where $A(\zeta)$ is positive, $\psi(\zeta)$ is real, and the $\overline{\gamma_\ell}(\zeta)$ are the point zeros corresponding to a given value of ζ. If ζ is forced to be of unit magnitude then the point zeros $\{\overline{\gamma_\ell}(\zeta)\}$ can be used to recover the the Fourier transform along a strip parallel to the (real) v-axis, in the manner described in §4.2.

If the value of ζ is now varied it is apparent that there is a shift in the positions of the point zeros $\{\overline{\gamma_\ell}(\zeta)\}$. Since $\mathcal{F}(\zeta,\gamma)$ is analytic the point zeros $\{\overline{\gamma_\ell}(\zeta)\}$ must migrate continuously with ζ. $\mathcal{F}(\zeta,\gamma)$ would cease to be analytic if any zero was to follow a discontinuous path. Since the complex parameter ζ has two real degrees of freedom, it is apparent that as ζ is varied across a two-dimensional surface of finite area, each individual point zero $\overline{\gamma_\ell}(\zeta)$ must also describe a two-dimensional surface. Hence the analytically continued Fourier transform of a two-dimensional image exists in a four-dimensional space, and is zero on a two-dimensional dimensional surface located within the space. Thus $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$ forms a continuous analytic surface called a zero-sheet (Lane et al. 1987) or in the terminology of Manolitsakis (1982) an analytic set (see also Sanz and Huang 1983b).

It is possible to describe $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$ by a multi-valued complex function of the form (Titchmarsh 1932)

$$\Omega(\zeta) = \gamma \tag{4.24}$$



(a) (b)

Figure 4.11: The relationship between (a) a ζ-contour and (b) the corresponding γ-contours.

In order to represent the behaviour of $\Omega(\zeta)$ as a single-valued function it is necessary to introduce branch cuts and branch points in the complex ζ-plane (Titchmarsh 1932). This is equivalent to replacing the complex ζ-plane with a Riemann surface. Whilst the above argument establishes the existence of a two-dimensional zero-sheet, it gives no idea of what a zero-sheet looks like. In order to examine this it is necessary to consider the behaviour of the zeros as ζ is varied continuously around a closed continuous path (henceforth called a contour) in the complex ζ-plane. Due to the analytic nature of $\mathcal{F}(\zeta, \gamma)$ any contour in the ζ-plane must give rise to continuous contours in the γ-plane, although if the ζ-contour crosses a branch-cut of $\Omega(\zeta)$ it may have to be traversed several times to fully define a particular γ-contour.

Consider now one of the point zeros associated with $\{\overline{\gamma_\ell}(\zeta_s)\}$, the starting point of the ζ-contour. If the ζ-contour does not intersect a branch cut of (4.24) then this point zero completely describes a γ-contour. If, however, the ζ-contour crosses a branch cut of (4.24) the path taken by a single point zero is not closed but must terminate upon another of the point zeros $\{\overline{\gamma_\ell}(\zeta_s)\}$. Only when the ζ-contour is traversed repeatedly does the path of the point zero return to its initial value.

Alternatively it is possible to follow the paths formed by all the $\{\overline{\gamma_\ell}(\zeta_s)\}$ for a single traverse of the given ζ-contour, Fig 4.11. In this case all these paths form at most $(M-1)$ γ-contours, when the ζ-contour does not cross a branch cut of $\Omega(\zeta)$.

4.6 Display of zero sheets

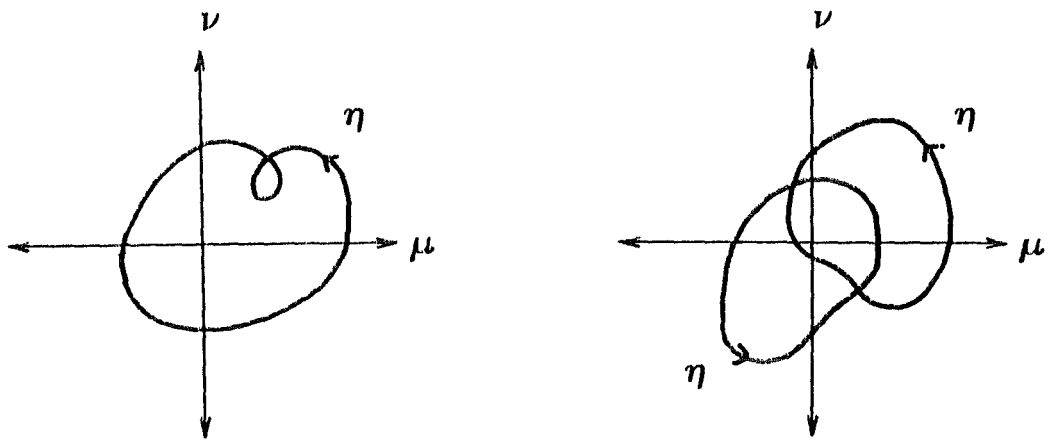
The display of two-dimensional surfaces in a four-dimensional space is a by no means easy task. In the following displays, the ζ-contour is defined for convenience by

$$\zeta = \rho e^{i\eta} \tag{4.25}$$

where ρ is fixed and η is varied between 0 and 2π. Varying ρ from 0 to ∞ forms a set of ζ-contours that fills the entire ζ-plane. In order to emphasise the four-dimensional nature of $\mathcal{F}(\zeta, \gamma)$ it is convenient to rewrite γ as

$$\gamma = \mu + i\nu \tag{4.26}$$

The (ζ, γ)-space, where ζ and γ are complex, is thus spanned by (ρ, η, μ, ν) where ρ,η,μ and ν are all real. The γ-contours formed for a particular value of ρ can be viewed as "slices" of the two-dimensional zero-sheet and can be stacked to form a two-dimensional surface in three-dimensional (ρ, μ, ν)-space. This space is projected from the full (ρ, η, μ, ν)-space by the expedient of drawing the γ-contours with η as a parameter as in Fig 4.12.



(a) (b)

Figure 4.12: Apparent self intersection of γ-contours due their parametric description as functions of η (a) single γ-contour (b) multiple γ-contours

Before displaying the γ-contours of an actual image it is important to understand some of the idiosyncracies introduced by this method of display. Fig 4.12a depicts an idealised γ-contour which is self-intersecting in (ρ, μ, ν)-space. Because the curve is drawn as a parameteric function of η this does not correspond to an intersection in (ρ, η, μ, ν)-space. Similarly the two separate γ-contours which appear to intersect in Fig 4.12b do not in general intersect in the unprojected (ρ, η, μ, ν)-space.

It is also of interest to consider the behaviour of the γ-contours as ρ tends to very large or small values. For example as $\rho \rightarrow 0$ there is an increasing dominance of the terms

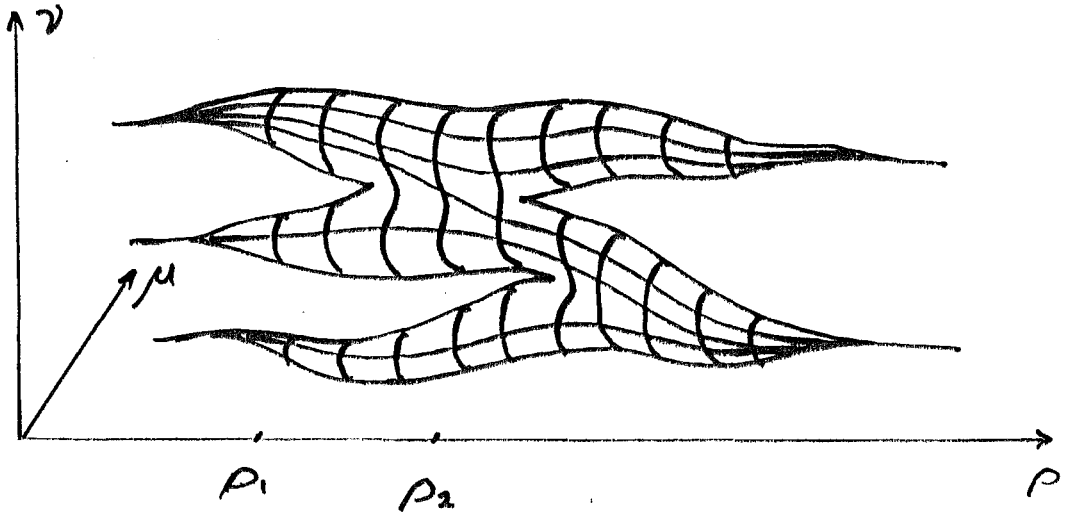


Figure 4.13: Idealised zero-sheet shown in (ρ, μ, ν)-space

for which $m = 0$ in (4.10). In the limit the two-dimensional polynomial, given by (4.10), effectively reduces to the one-dimensional polynomial

$$\mathcal{F}(\gamma) = K \sum_{m=0}^{M-1} f_{0,n} \gamma^m \tag{4.27}$$

where K is a constant. The corresponding γ-contours thus tend to a set of discrete points determined by the pixels on one edge of $B_f(x, y)$. As $\rho \to \infty$ the γ-contours are fixed by the the pixels on the opposite edge of $B_f(x, y)$.

Fig 4.13 shows a hypothetical zero-sheet in (ρ, μ, ν)-space corresponding to a 4 x 4 image. Although considerably simpler than the zero-sheet which would typically be obtained from an actual 4 x 4 image it illustrates some of the problems encountered in mapping a zero-sheet. The point zeros $\{\overline{\gamma_\ell}(\zeta)\}$ corresponding to a particular value of ζ lie on the zero-sheet.

It is possible to obtain the γ-contours for any value of ρ by finding the intersection of the zero-sheet with the plane $\rho = C$, where C is a real positive constant, as shown in Fig 4.14. Because it is not possible to choose ρ so that there is only one γ-contour, it is apparent that no single value of ρ is sufficient to show that the zero-sheet consists of single surface. In order to find trajectories along the zero-sheet between arbitrary points on its surface it is necessary to introduce the concept of ζ_ρ-paths, which are similar to ζ-contours but η is fixed and ρ is varied instead of vice versa. The term path instead of contour is employed because a path although continuous need not be closed. Fig 4.15 illustrates schematically how appropriate use of ζ-contours and ζ_ρ-paths can be used to connect the $\{\overline{\gamma_\ell}(\zeta_s)\}$.

The preceding discussion has described the mapping of zero-sheets in theory.

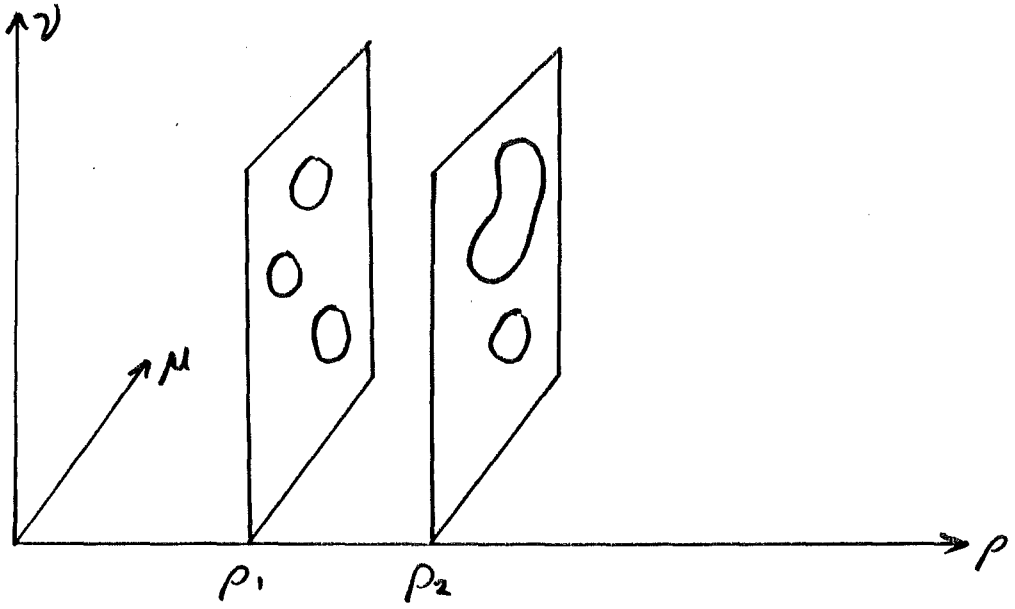
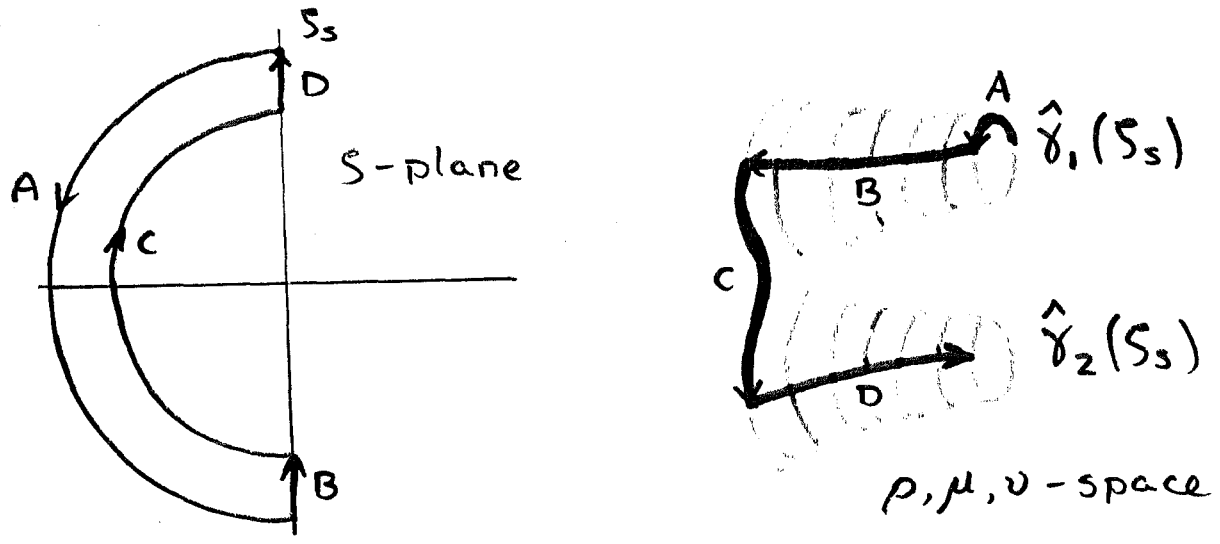


Figure 4.14: Intersection of a planes define by $\rho = \rho_n$ with the zero-sheet shown in Fig 4.13



(a) (b)

Figure 4.15: Schematic of the use of zero-paths and zero-contours to connect $\{\overline{\gamma_\ell(\zeta)}\}$ (a) ζ-contours and ζ_ρ-paths (b) Equivalent γ-contours and γ_ρ-paths.

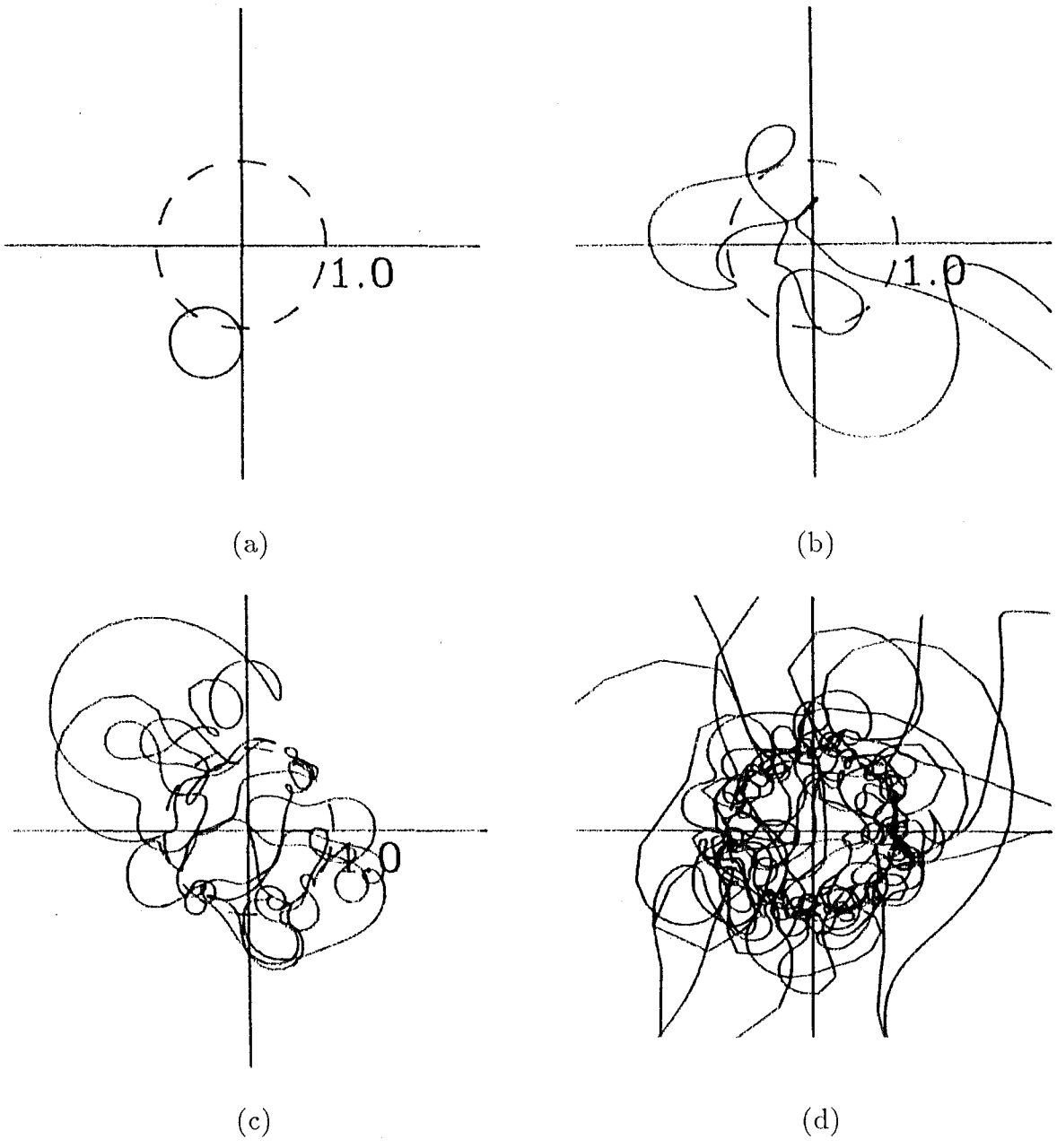
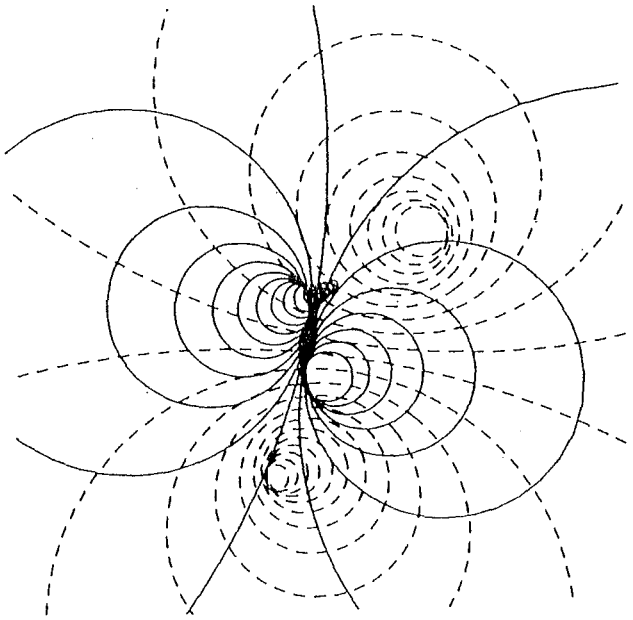


Figure 4.16: Actual γ-contours corresponding to the ζ-contour formed by $\rho = 1.0$. The images are arrays of pseudo-random complex numbers (a) 2 x 2 pixel image (b) 4 x 4 pixel image (c) 8 x 8 pixel image (d) 16 x 16 pixel image.



SEE ERRATA

Figure 4.17: Zero-sheet computed from the convolution of the pixellated images given in Tables 4.1 and 4.2

In practice, as always, the situation is more complicated. Firstly, visualising the four-dimensional space by a projection into three-dimensional space becomes less practical as the intricacy of the γ-contours increases. Fig 4.16 shows examples of actual γ-contours for images formed of random complex numbers.

4.7 Deconvolution of two dimensional polynomials

The previous section discusses why the Z-transform of a two dimensional pixellated object is zero on a continuous two-dimensional surface in a four-dimensional space. In Z-space a convolution can also be represented by a product (Oppenheim and Schafer 1975)

$$\mathcal{G}(\zeta,\gamma) = \mathcal{F}(\zeta,\gamma)\mathcal{H}(\zeta,\gamma) \tag{4.28}$$

Consequently $\mathcal{G}(\zeta,\gamma)$ is equal to zero when either $\mathcal{F}(\zeta,\gamma)$ or $\mathcal{H}(\zeta,\gamma)$ is equal to zero. Thus $\mathcal{Z}[\mathcal{G}(\zeta,\gamma)]$ must be the union of $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$ and $\mathcal{Z}[\mathcal{H}(\zeta,\gamma)]$

$$\mathcal{Z}[\mathcal{G}(\zeta,\gamma)] = \mathcal{Z}[\mathcal{F}(\zeta,\gamma)] \cup \mathcal{Z}[\mathcal{H}(\zeta,\gamma)] \tag{4.29}$$

Since it is possible to recover an image from its zero-sheet, a topic dealt with in §4.8, partitioning the zero-sheet $\mathcal{Z}[\mathcal{G}(\zeta,\gamma)]$ into $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$ and $\mathcal{Z}[\mathcal{H}(\zeta,\gamma)]$ is equivalent to blindly deconvolving $g(x,y)$. The zero-sheet of the the convolution of the images given in Tables 4.1 and 4.2 is shown in Fig 4.17. The zero-sheet shown in Fig 4.17 can be separated into two distinct analytic surfaces, as shown in Figs 4.18a and 4.18b. The images recovered from the zero-sheets of Figs 4.18a and b are, to within a complex scale factor, the images whose pixel values are listed in Tables 4.1 and 4.2.

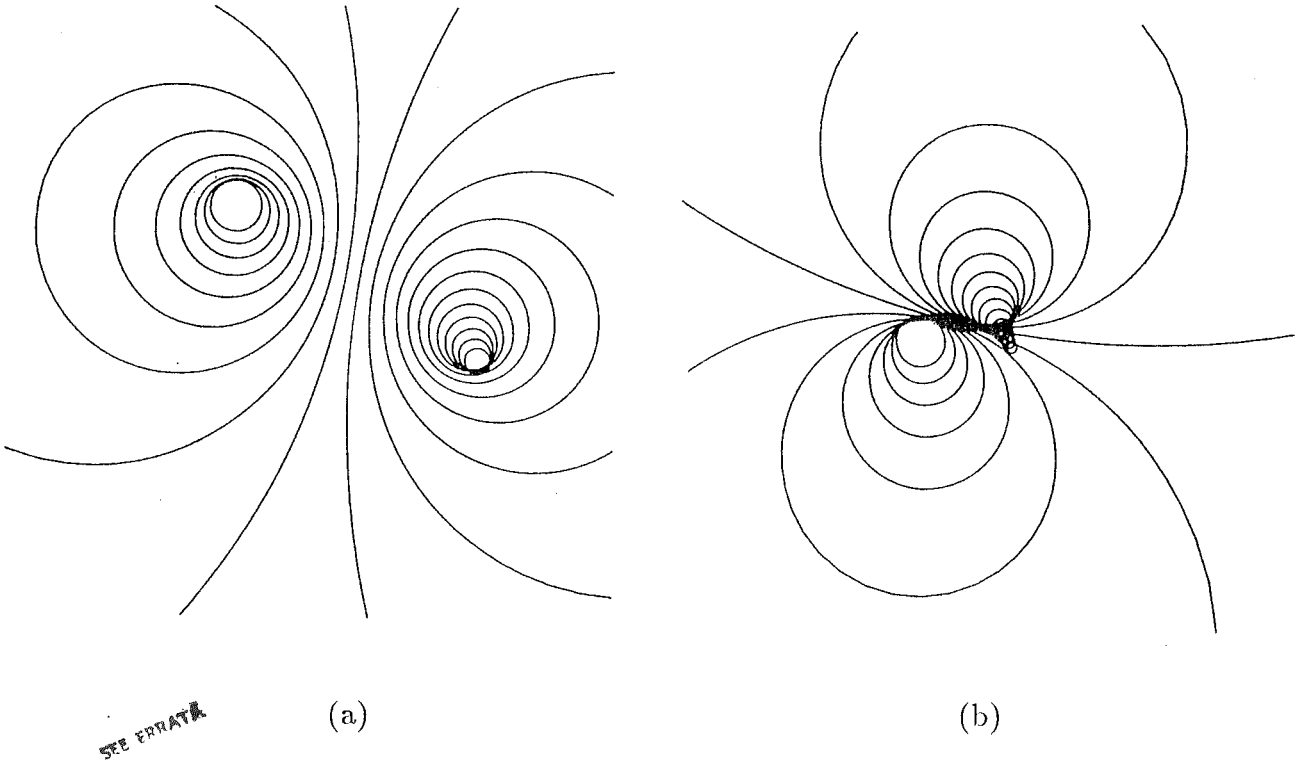


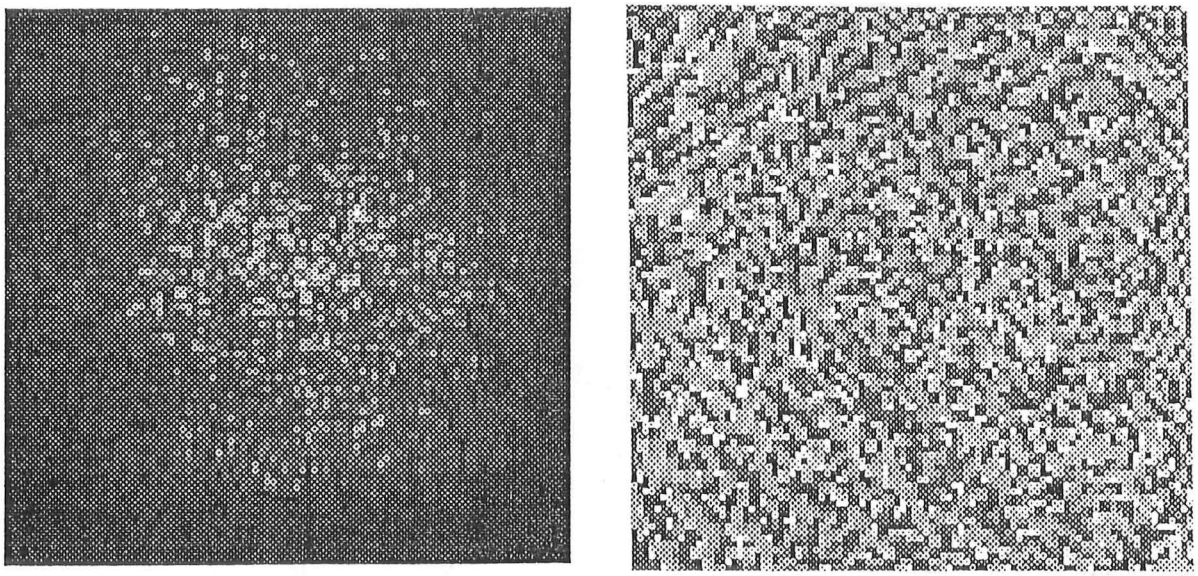
Figure 4.18: Zero-sheets corresponding to the pixellated images given in Table 4.1 (a), and Table 4.2 (b)

	y	
x	0	1
0	$1.31 + i2.18$	$1.48 + i0.43$
1	$3.19 + i2.56$	$-0.64 + i2.05$

Table 4.1: Complex amplitudes and x, y coordinates of pixels defining a particular 2 x 2 pixel image.

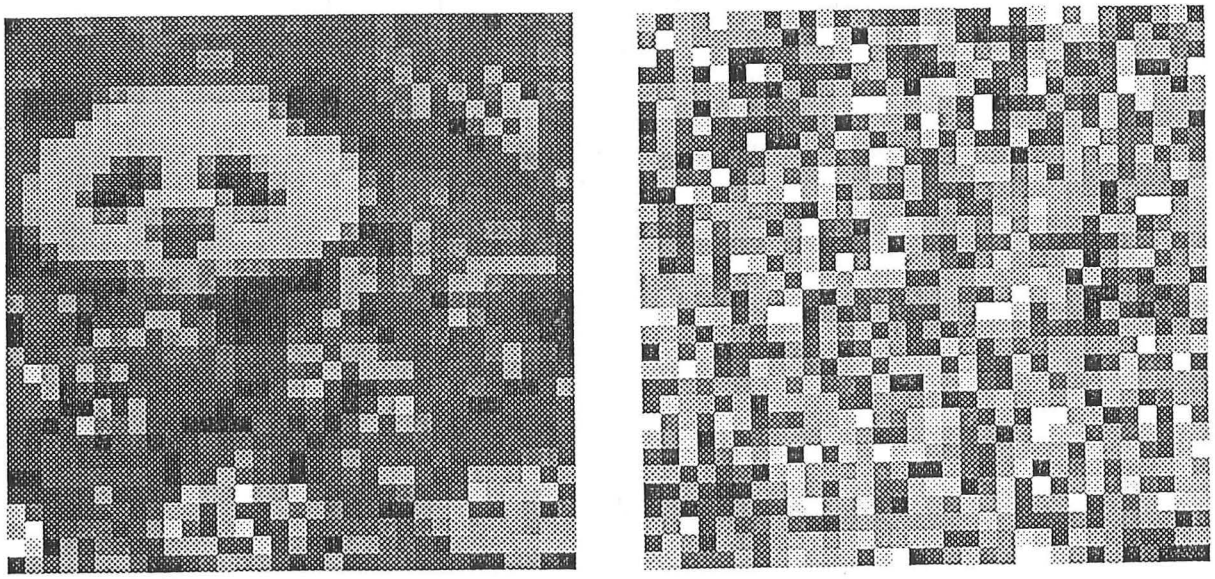
	y	
x	0	1
0	$-0.64 - i2.05$	$3.19 - i2.56$
1	$1.48 - i0.43$	$1.31 - i2.18$

Table 4.2: Complex amplitudes and x, y coordinates of pixels defining another 2 x 2 pixel image.



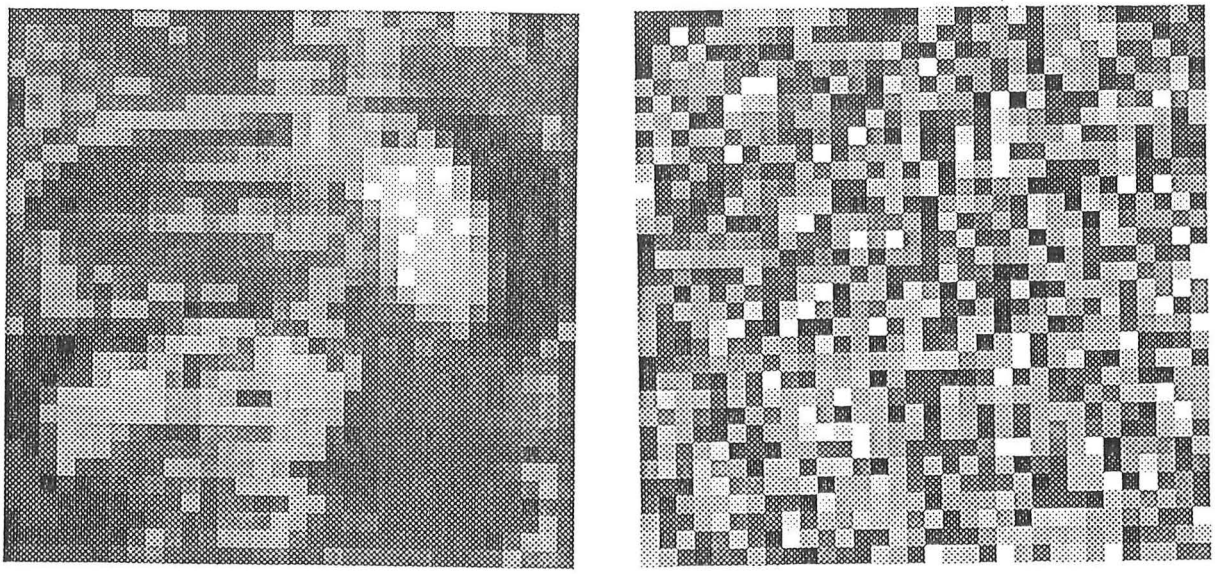
(a) (b)

Figure 4.19: 63 x 63 pixel complex convolution of the two images shown in Figs 4.20 and 4.21 (a) magnitude and (b) phase



(a) (b)

Figure 4.20: Complex 32 x 32 pixel image



(a) (b)

Figure 4.21: Complex 32 x 32 pixel image (a) magnitude and (b) phase.

One major advantage of zero-sheet based deconvolution is that it does not require the image to be positive, or even real. In the absence of a priori information, however, it is impossible to determine which of the two components is in fact the image and which is the psf. Fortunately, this ambiguity can usually be resolved by a further examination of the physical process involved. In general, the number of zero-sheets in a convolution is equal to the number of components in the convolution.

Using the principle of zero-sheet separation it is possible, in the absence of noise, to deconvolve quite large convolutions. Fig 4.19 shows a 63 x 63 pixel complex convolution (i.e. each pixel has both magnitude and phase). This was successfully deconvolved to the two complex 32 x 32 pixel objects shown in Figs 4.20 and 4.21. The technique is limited both by the sophistication of the zero-sheet mapping algorithm and the size of polynomial which can be factored. In order to map the zero sheet of an M x M image it is necessary to find reliably the zeros of many M^{th} order polynomials. Consequently, mapping a zero-sheet requires a very robust algorithm, since any inaccuracies in zero location preclude successful deconvolution.

The algorithm starts by using the standard algorithm CPOLY (Jenkins and Traub 1971) to find the initial point zeros at the start of a zero contour or path (cf §4.5). Rather than employ a general purpose zero finding routine for subsequent calculations of zero positions, a specialised routine was written in order to take advantage of the analytic nature of the zero-sheet. This is possible because there is little difference between the point zeros $\{\overline{\gamma_\ell}(\zeta_0)\}$ and $\{\overline{\gamma_\ell}(\zeta_0 + \Delta\zeta)\}$ when $\Delta\zeta$ is small. A more sophisticated approach relies on predicting the zero positions at the the next increment in ζ by using the last two sets of zero positions,

$$\hat{\gamma}(\zeta_\ell) = 2\overline{\gamma_\ell}(\zeta - \Delta\zeta) - \overline{\gamma_\ell}(\zeta - 2\Delta\zeta) \; for \; \ell = 0, \ldots, M-1 \tag{4.30}$$

where $\hat{\gamma}_\ell$ denotes an estimate of the true zero positions. Each $\hat{\gamma}_\ell$ is then refined by using

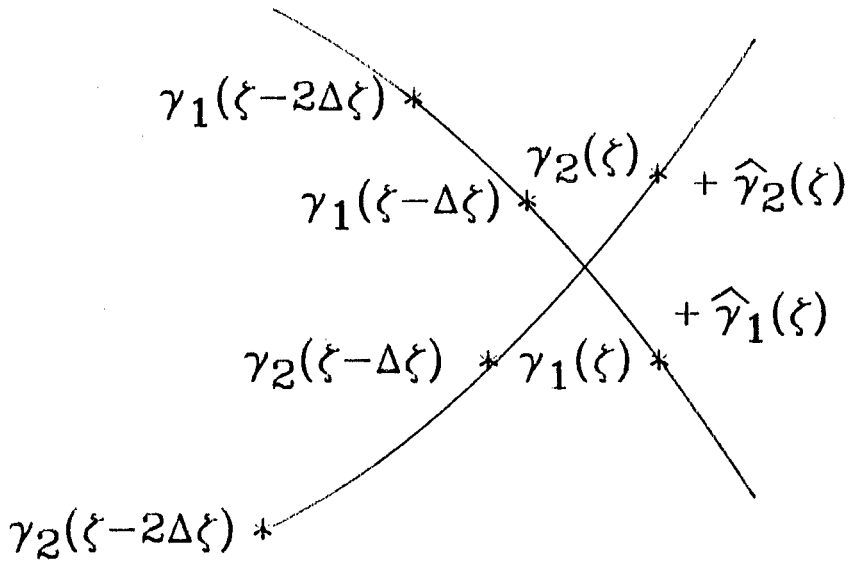


Figure 4.22: Application of continuity of the first derivative to separate intersecting zero-sheets.

a Newton-Raphson search.

$$\hat{\gamma_\ell}^{n+1} = \hat{\gamma_\ell}^n - \frac{\mathcal{F}(\zeta, \hat{\gamma_\ell}^n)}{\frac{\partial \mathcal{F}'(\zeta, \hat{\gamma_\ell}^n)}{\partial \gamma}} \tag{4.31}$$

until $\hat{\gamma_\ell}^n \rightarrow \overline{\gamma_\ell}$.

Using zero positions predicted from previous zero positions also helps avoid confusion between zero contours which are very close in Z-space, because it helps enforce continuity of the first derivative of $\mathcal{Z}[\mathcal{F}(\zeta, \gamma)]$. An example where the continuity of the first derivative must be used to determine the correct zero contours is when two zero-contours intersect as shown in Fig 4.22. Since the zero-sheet is analytic both the zero-sheet and its first derivative must be continuous and it is then possible to unambiguously separate the two intersecting zero-sheets.

Some care must also be taken in designing the zero location procedure for a fixed value of ζ even when predicted zero positions are available. Using $\mathcal{F}_{\zeta_0}(\gamma)$, as defined in (4.14), it is convenient to introduce $M_r[\mathcal{F}_{\zeta_0}(\gamma)]$ which is defined to be the largest value of $|\mathcal{F}_{\zeta_0}(\gamma)|$ when $|\gamma| = r$. The maximum modulus theorem (Wilkinson 1963) states that $M_r[\mathcal{F}_{\zeta_0}(\gamma)]$ is a monotonically increasing function of r. This rate of increase is dramatic for $r > 1$. A problem which arises when locating zeros of large modulus is that small errors in the predicted values can result when numbers appearing in (4.31) are larger than those representable on a typical digital computer (commonly referred to as numeric overflow). Furthermore, the predicted zero positions of zeros of large modulus are also less likely to be accurate, further increasing the chances of overflow in (4.31).

In order to overcome the problems of numeric overflow two measures were employed. The first was to detect numeric overflow when it occurred. After an overflow was encountered the zero location procedure was terminated and restarted after the polynomial coefficients had all been divided by a large constant to reduce the value of $\mathcal{F}_{\zeta_0}(\gamma)$.

The second measure was to employ polynomial deflation. Once a zero of $\mathcal{F}_{\zeta_0}(\gamma)$ was successfully located the quotient

$$\mathcal{F}'_{\zeta_0}(\gamma) = \frac{\mathcal{F}_{\zeta_0}(\gamma)}{\gamma - \overline{\gamma_\ell}} \tag{4.32}$$

was formed and the quotient (or deflated polynomial) used to locate subsequent zeros. Thus as more zeros are found the "deflated" polynomial becomes of lower order and consequently less prone to numeric overflow. The process of deflation is numerically well conditioned provided the zeros are found in order of increasing modulus (Wilkinson 1963).

Although it is possible to map the entire zero-sheet using the above technique, in general it is only necessary to determine the zero-sheet at a number of specific points in Fourier space. Exactly which, and how many points need to be determined is a function of the technique being employed to recover the image-form. This is discussed further in the next section.

4.8 Image recovery from zero-sheets

The problem of recovering a function from the points where it is zero has been the subject of much ongoing interest in recent years (Sanz 1985b, Curtis and Oppenheim 1987, Curtis et al. 1985, Zakhor and Izraelevitz 1986, Rotem and Zeevi 1986). In one dimension it is relatively easy to recover a compact image from a set of zeros (§3.1). In two dimensions, however, the problem is complicated by the fact that there are an infinite number of points on the zero-sheet.

There are two major approaches to recovering an image from a zero-sheet. Firstly, it is possible to use one-dimensional projections in order to determine the Fourier transform along lines in Fourier space. When these lines are made to correspond to a sampling grid spaced at the Nyquist frequency it is then possible to retrieve the image using the inverse discrete Fourier transform. Secondly, there is a linear equations approach, as introduced by Curtis et al (1985). It should be noted in passing that a third technique, based on an iterative loop, also exists (Curtis et al. 1985, Sayegh et al 1987). This technique usually requires a good estimate of the starting image in order to be successful and so is not discussed further here.

4.8.1 Use of one-dimensional projections

In order to employ one-dimensional projections to recover an image from its zero-sheet it is necessary to find $\{\overline{\gamma_\ell}(\zeta_j)\}$ where ζ_j are the points given by

$$\zeta_j = e^{\frac{i2\pi m}{M}} \; for \; m = 0, 1, \ldots, M-1 \tag{4.33}$$

Note from (4.13) that the ζ_j correspond to the values of u at which $F(u,v)$ is evaluated using the DFT (§1.7).

By substituting ζ_j and $\overline{\gamma_\ell}(\zeta_j)$ in (4.23) and evaluating at $\gamma = e^{\frac{i2\pi m}{M}}$ for $k = 0, 1, \ldots, M-1$ it is possible to reconstruct all the values of $F(u,v)$ needed to evaluate the image, to within a complex scale factor, by the inverse DFT. Before this is possible, however, it is necessary to evaluate $\psi(\zeta_j)$ and $A(\zeta_j)$ so that they can be inserted into (4.23). This is done by finding where the zero-sheet intersects the point $\gamma = 1.0$, thereby

generating the point zeros $\{\overline{\zeta_\ell}\}$ for $\ell = 1, 2, \ldots, M$ which characterise $\mathcal{F}(\zeta, 1)$. Since it is only possible to recover an image from its zero-sheet to within a complex constant, it is possible to define

$$\mathcal{F}(\zeta, 1) = \prod_{\ell=1}^{M} (\zeta - \overline{\zeta_\ell}) \tag{4.34}$$

$\psi(\zeta_0)$ is then chosen such that (4.23) and (4.34) give the same value for $\mathcal{F}(1,1)$. The remainder of the $\psi(\zeta_j)$ are found by comparing the values provided by (4.23) and (4.34).

4.8.2 Linear equations approach

The linear equations approach has recently been the subject of a review by Curtis and Oppenheim (1987). Although this deals primarily with image recovery from the zero crossings of an image, the technique was applied to image recovery from the equivalent of a zero contour by Izraelevitz and Lim (1987). The technique relies on formulating a set of linear equations with the pixels of the image as unknowns. To illustrate this process, consider N point zeros (ζ_j, γ_j), for j = 1,2,...,N, zeros lying on a zero-sheet. For each zero

$$\mathcal{F}(\zeta_j, \gamma_j) = \sum_{m,n=0}^{M-1} f_{m,n} \zeta_j^m \gamma_j^n = 0 \text{ for } j = 1, \ldots, N \tag{4.35}$$

Reformulating (4.35) as a linear equation with the $f_{m,n}$ as unknowns yields,

$$\sum_{m,n=1}^{M-1} (\zeta_j^m \gamma_j^n) f_{m,n} = 0 \text{ for } j = 1, \ldots, N \tag{4.36}$$

In order to avoid the trivial solution $f_{m,n} = 0 \ \forall \ m, n$ it is necessary to force one of the pixels of f to be nonzero, e.g.by setting $f_{0,0} = 1$. The solution of this system of linear equations results in, to within a scale factor, the image corresponding to the zero-sheet.

One of the major difficulties with using linear equations is determining how many samples of the zero-sheet are necessary to uniquely determine the image-form. The approach taken by Curtis et al. (1985) relies upon determining the maximum number of points two polynomials can have in common without sharing a common factor.

As an example consider the zero sheets of two arbitrary two-dimensional polynomials defined by

$$K(\zeta, \gamma) = \sum_{m,n=0}^{M-1} k_{m,n} \zeta^m \gamma^n = 0 \tag{4.37}$$

and

$$L(\zeta, \gamma) = \sum_{m,n=0}^{M-1} l_{m,n} \zeta^m \gamma^n = 0 \tag{4.38}$$

If $\mathcal{Z}[K(\zeta, \gamma)]$ and $\mathcal{Z}[L(\zeta, \gamma)]$ can have at most J points in common, then J samples of the zero-sheet should be sufficient to recover the image-form of either $k(x, y)$ or $l(x, y)$.

§4.5 shows that $\mathcal{Z}[K(\zeta, \gamma)]$ and $\mathcal{Z}[L(\zeta, \gamma)]$ are two-dimensional surfaces existing in a four-dimensional space. In general, two M-dimensional surfaces existing in an N-dimensional space intersect in a space of $(2M - N)$ (Walker 1950, p40). Hence $\mathcal{Z}[K(\zeta, \gamma)]$ and $\mathcal{Z}[L(\zeta, \gamma)]$ would be expected to intersect in a zero-dimensional space, i.e. at discrete points.

In order to determine the points of intersection of $\mathcal{Z}[K(\zeta,\gamma)]$ and $\mathcal{Z}[L(\zeta,\gamma)]$ it is convenient to rewrite (4.38) and (4.39) as

$$K(\zeta,\gamma) = \sum_{m,n=0}^{M-1} \overline{k_m(\gamma)}\zeta^n \tag{4.39}$$

and

$$L(\zeta,\gamma) = \sum_{m,n=0}^{M-1} \overline{l_m(\gamma)}\zeta^n \tag{4.40}$$

where each $\overline{k_m}(\gamma)$ and $\overline{l_m(\gamma)}$ is a polynomial of order $(M-1)$ in ζ. The resultant polynomial is defined to be the determinant (Zakhor and Izraelevitz 1986)

$$\begin{vmatrix}
\overline{k_0(\gamma)} & \overline{k_1(\gamma)} & \cdots & \cdots & \overline{k_{M-1}(\gamma)} & 0 & \cdots & 0 \\
0 & \overline{k_0(\gamma)} & \overline{k_1(\gamma)} & \cdots & \overline{k_{M-2}(\gamma)} & \overline{k_{M-1}(\gamma)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & 0 \\
0 & 0 & \cdots & \overline{k_0(\gamma)} & \cdots & \cdots & \cdots & \overline{k_{M-1}(\gamma)} \\
\overline{l_0(\gamma)} & \overline{l_1(\gamma)} & \cdots & \cdots & \overline{l_{M-1}(\gamma)} & 0 & \cdots & 0 \\
0 & \overline{l_0(\gamma)} & \overline{l_1(\gamma)} & \cdots & \overline{l_{M-2}(\gamma)} & \overline{l_{M-1}(\gamma)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & 0 \\
0 & 0 & \cdots & \overline{l_0(\gamma)} & \cdots & \cdots & \cdots & \overline{l_{M-1}(\gamma)}
\end{vmatrix} \tag{4.41}$$

which is seen to be a polynomial in γ. Since the zeros of the resultant polynomial give the points of intersection of $\mathcal{Z}[K(\zeta,\gamma)]$ and $\mathcal{Z}[L(\zeta,\gamma)]$, the number of points of intersection is given by the order of the resultant polynomial (Walker 1950). Thus $\mathcal{Z}[K(\zeta,\gamma)]$ and $\mathcal{Z}[L(\zeta,\gamma)]$ can be expected to intersect at $2(M-1)^2$ points.

Implicit in the above derivation is knowledge of the order of $\mathcal{Z}[K(\zeta,\gamma)]$ and $\mathcal{Z}[L(\zeta,\gamma)]$. Since the number of points of intersection is a function of the assumed order M, increasing the assumed order increases the number of samples required to effect recovery of the image-form. As noted by Sanz (1985b) it is not possible, in general, to uniquely recover an image from samples of its zero-sheet, since some images are convolutions of component images. This is readily apparent from (4.29) because $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$ is a subset of a $\mathcal{Z}[\mathcal{G}(\zeta,\gamma)]$. Thus when the true image is a convolution it is important to ensure that there are sufficient sample points on the zero-sheet to permit each individual component visibility to be separately recovered.

In practice however, provided there are more samples of the zero-sheet than there are pixels in the image (Curtis and Oppenheim 1987) an image-form can be recovered. The sensitivity of the recovery procedure to error can be reduced by increasing the number of samples taken of the zero-sheet. It should be noted that in order to ensure the optimal image recovery, some form of regularisation such as a weighted least squares technique (Taylor 1982) should be employed because, as is noted in §6.4, not all portions of a noisy zero-sheet are known to an equal degree of accuracy.

4.9 Effects of noise on zero-sheets

§4.3 notes that a noisy convolution can no longer be deconvolved into smaller images. It is of interest to consider this in the context of zero-sheets. Since the zero-sheet of an

irreducible image must be continuous, the addition of noise must cause $\mathcal{Z}[\mathcal{F}(\zeta,\gamma)]$ and $\mathcal{Z}[H(\zeta,\gamma)]$ to become linked as part of a single zero-sheet.

In general the displacement of a zero-sheet is proportional to the amount of noise added. Consequently in order for an infinitesimal amount of noise to cause the linkage of two zero-sheets, they must intersect in a subspace of the the space spanned by the zero-sheet. In fact, as indicated in §4.8.2, the zero-sheets of two-dimensional visibilities intersect at discrete points.

Fig 4.23 illustrates how of the zero-sheet of the uncontaminated convolution shown in Fig 4.17 becomes increasingly distorted with the addition of noise. The level of contamination ϱ is defined by

$$\varrho = \frac{\int_{(x,y)} |c(x,y)|^2}{\int_{(x,y)} |g(x,y)|^2}, \tag{4.42}$$

where $g(x,y)$ and $c(x,y)$ are defined in (2.1).

At $\varrho = 10^{-4}$ a single "bridge" or linkage between the zero-sheets is apparent and is marked with a triangle in Fig 4.23a. The distortion of the zero-sheet becomes more pronounced in Figs 4.23b and c until in Fig 4.23d the zero-sheets comprising the original convolution are unrecognisable. It should be noted that Fig 4.23d corresponds to a very high contamination level, in that the pixels of the original and noisy convolutions differ by as much as 30%.

Fig 4.24 shows, for a 5 x 5 pixel image convolution, the γ-contours corresponding to the ζ-contour defined by $\rho = 1$ (§4.6). An intersection of $\mathcal{Z}[F(\zeta,\gamma)]$ and $\mathcal{Z}[H(\zeta,\gamma)]$ occurs within the small square drawn in Fig 4.24a and is enlarged for clarity in Fig 4.24b. The addition of noise transforms this point of intersection into a pair of close but separated sharp bends (Fig 4.24c). The zero-sheets are now no longer distinct but have been "bridged" to form a single analytic surface. Increasing the level of noise causes these bridges to become less distinct, as illustrated in Fig 4.24d.

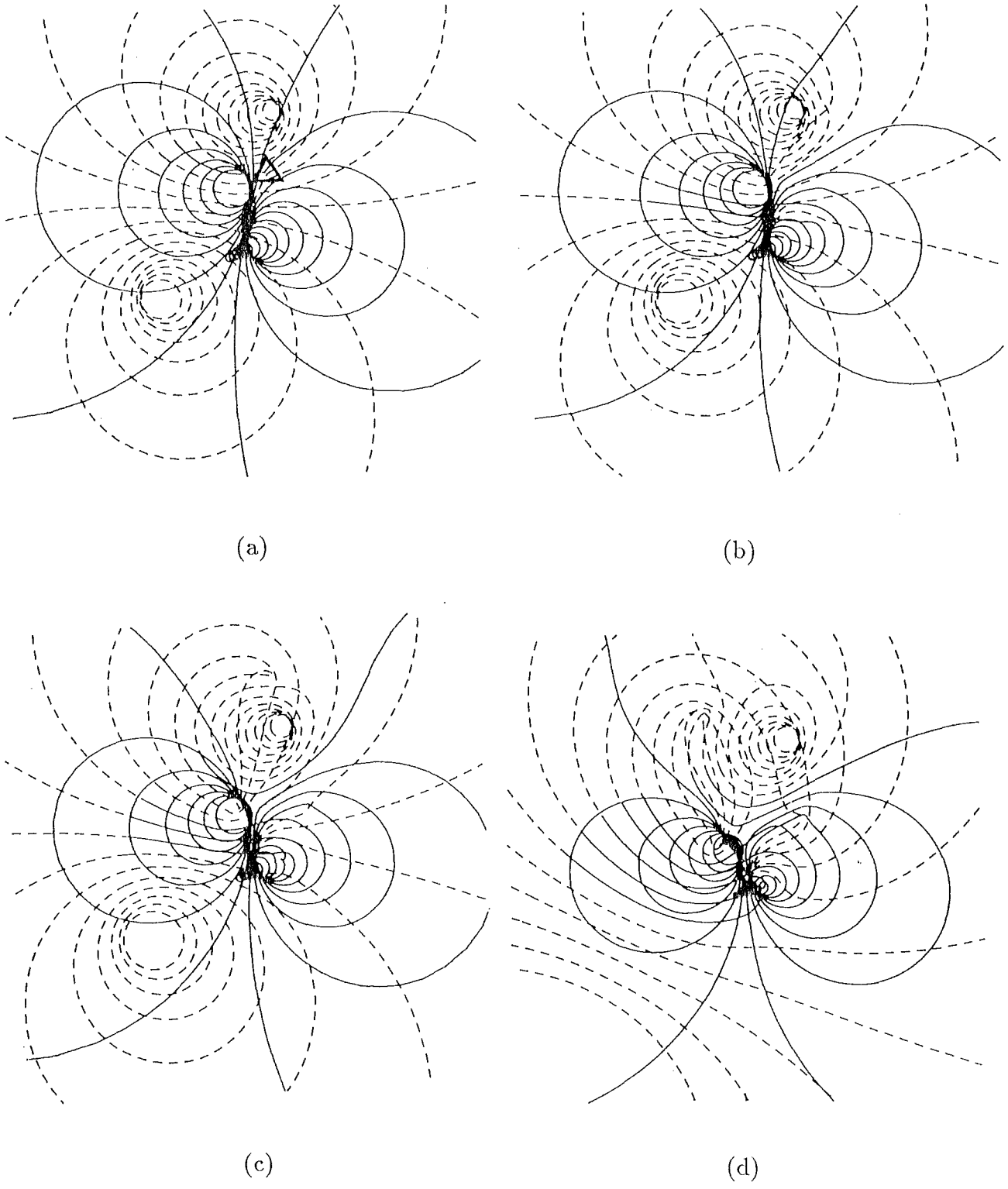
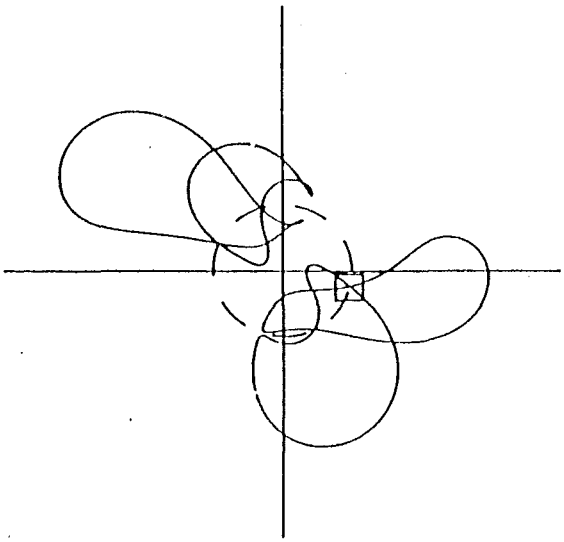
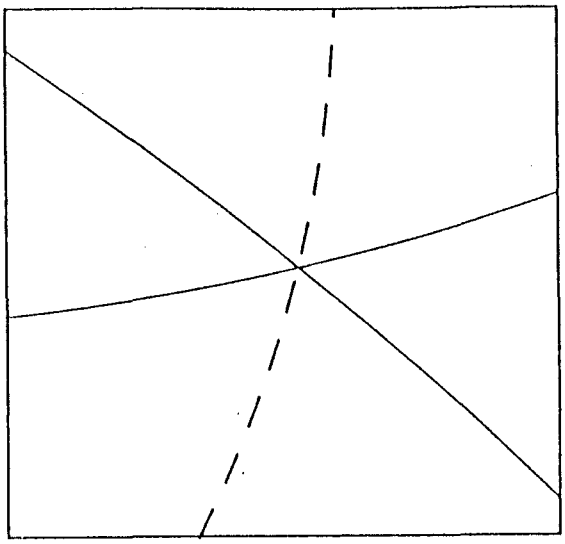


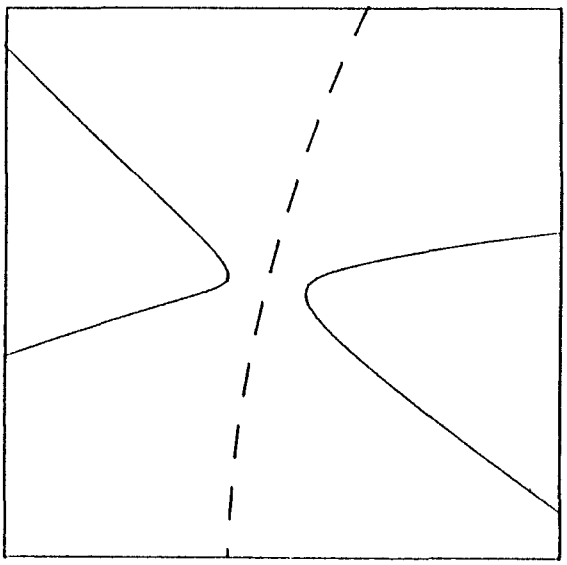
Figure 4.23: The effects of noise on the zero-sheet shown in 4.17. (a) $\varrho = 10^{-4}$ (b) $\varrho = 10^{-3}$ (c) $\varrho = 10^{-2}$ (d) $\varrho = 10^{-1}$.



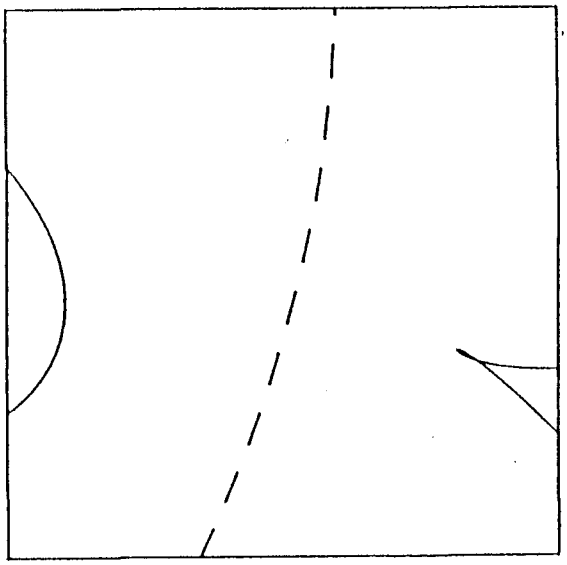
(a)



(b)



(c)



(d)

Figure 4.24: Zero-contours of a 5 x 5 pixel convolution corresponding to $\rho = 1.0$. (a) The complex zero-contours (b) enlargement of the zero-contours within the small square in Fig 4.24a (c) Same part of the zero-contours when noise is added (d) Same part of the zero-contours with the noise 100 times greater.

Chapter 5

ITERATIVE PROCESSING

Phase retrieval has recently been the subject of a number of excellent reviews (Mnyama 1987; Bates and Mnyama 1986; Fienup and Dainty 1986). As these works, and the references therein, already provide comprehensive surveys of available phase retrieval techniques, this chapter concentrates on evaluating the most practical alternative to analytic solution of the phase problem, namely the iterative algorithms proposed by Fienup (Fienup 1987; Fright 1984; Fienup 1982).

The major difficulty in solving the Fourier phase problem is that the available information is distributed between Fourier and image space. Thus although an estimate of the image may have the correct support in image space and may meet other image space constraints such as positivity, there is no guarantee that it has the correct Fourier magnitude. If the estimate is Fourier transformed and made to comply with the constraints in Fourier space it will, in general, no longer be of the correct support when transformed back into image space. The difficulties in simultaneously meeting constraints in both Fourier and image space are further compounded by the global nature of the Fourier transform (cf §1.1), which means that the value of $F(u,v)$ at any point in Fourier space is a function of the value of $f(x,y)$ at all points in image space and vice versa. It is thus necessary to iterate between image and Fourier space to find a "feasible" solution (Trussell and Civanlar 1984) which simultaneously meets the constraints in the separate domains.

Iterative image recovery, rather than attempting to find an exact solution, views the restoration problem as one of constrained optimisation. An estimate of the true image is constrained in image (Fourier) space. The constrained estimate is Fourier transformed and a numerical measure of the deviation from the known constraints in Fourier (image) space is then formed. Iterative phase retrieval aims to minimise this numerical deviation, or error metric, in Fourier (image) space, whilst satisfying the imposed constraints in image (Fourier) space.

§5.1 introduces the basic iterative loop, one which has been employed in the Fourier phase problem (Fienup 1982), electron microscopy (Missell 1978), the magnitude problem (Hayes 1982) and bandlimited extrapolation (Papoulis 1984). The basic operation of the iterative loop is described. As with all forms of iterative processing there are two important points. Firstly there must be some way of measuring the error in the current estimate, i.e. how far it deviates from the true image. Secondly the loop must be configured so that this error decreases, or the estimate converges to the true solution.

The convergence of the basic iterative process has been the subject of much analysis, which can be divided into analysis based on projection onto convex sets (Youla 1978;

Tom et al. 1981; Youla and Webb 1982; Levi and Stark 1982; Hayes 1982) and analysis which views the iterative loop as a steepest descent search (Fienup 1982). Unfortunately both these analyses can only be applied when the constraints are applied in a relatively simple manner.

The accuracy of the constraints has, in general, a significant effect on the rate of convergence of the iterative algorithms. The most common example is estimation of the size of the object which is discussed in §5.2. There are two ways that the "strength" of the constraints affects convergence. Firstly the speed of convergence is usually increased when more constraints are applied. Conversely convergence as measured by deviation from the known constraints may appear, initially at least, to be faster when less constraints are applied. This is because any error based on summing the amount by which constraints are violated in image (Fourier) space must be reduced when less constraints are applied.

The discussion of §5.3 concentrates on the phase problem, with discussion of the magnitude problem deferred until chapter 6. Perhaps the most significant advance in phase retrieval in recent years has been Fienup's adaption of the basic iterative processing loop. The hybrid input-output algorithm as it is known relies on a non-linear application of the image space constraints in a manner described in §5.3. Originally the Fienup algorithms were only applied to positive images, but this has recently been extended firstly to images with specialised supports (Fienup 1987) and then to general complex images (Lane 1987). Although it is traditional to combine hybrid input-output with error reduction (Fienup and Wackerman 1986), it is demonstrated here, however, that this usually slows convergence to the true image.

Hybrid input-output, as the name suggests, is a combination of two different techniques. When combining the techniques a feedback parameter β is used in order to "urge" pixels violating the image space constraints to more acceptable values. As with all forms of feedback the exact level of β affects the performance of the algorithm. Selecting too small a value of β causes slow convergence, whilst too large a value of β results in instability of the feedback progress. The variation of convergence with β is illustrated in §5.4.

§5.5 is concerned with the effect of support size on the convergence of the Fienup iterations. The test images chosen consist of a sum of gaussians of varying amplitudes. These images are especially difficult to reconstruct because they are by definition of infinite extent in image space. Enforcing a finite support in image space thus necessarily involves an approximation. Three types of image are investigated, positive, bipolar and complex. The reconstructions with which §5.5 is illustrated include the first successful reconstructions of this class of bipolar and complex images.

The next section, §5.6, briefly discusses the effects of noise on image reconstruction. This subject is currently under comphrehensive investigation by McCallum at the University of Canterbury. The main purpose of this section is to show that the iterative recovery algorithms are robust in the presence of noise, i.e. the quality of reconstruction is roughly predictable by the quantity of noise added to the Fourier magnitude.

The final section deals with stagnation, or the failure of the iterative process to converge within a reasonable number of iterations. There are a number of ways the iterative process can stall at an estimate significantly different from the true image. The basic problem is that although there is only one image-form, there are still several interim solutions. For example although $f(x,y)$ and $f^*(-x,-y)$ correspond to the same image-form it is possible for the iterative loop to stagnate between these two possible manifestations

of the image-form. Fortunately stagnant solutions are usually readily recognisable. §5.7 discusses the identification of these stagnant solutions and proposes a new technique for accelerating convergence.

5.1 The basic iterative loop

An intuitively appealing method of finding a solution which satisfies constraints in both Fourier and image space is the iterative processing loop shown in Fig 5.1.

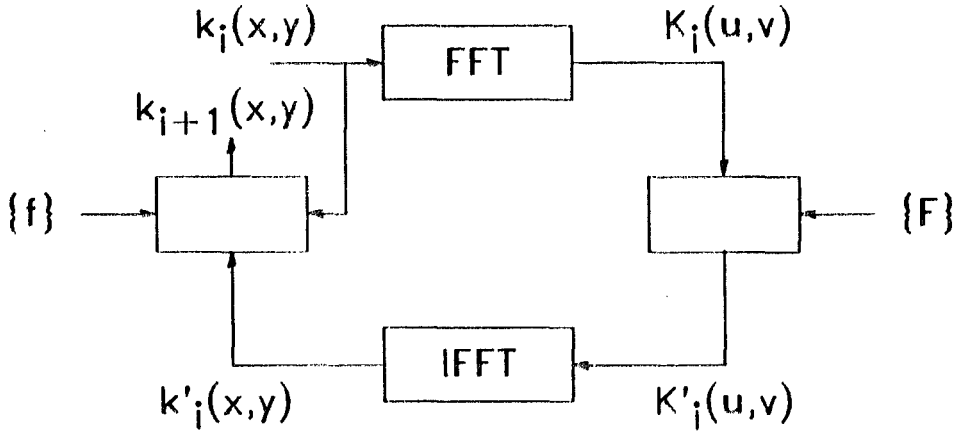


Figure 5.1: The basic iterative processing loop. Incomplete information available about an image $f(x,y)$ and its visibility $F(u,v)$ is denoted by $\{f\}$ and $\{F\}$ respectively.

In order to maintain notational consistency with previous work both the image and its visibility are referred to as functions of continuous variables, (x,y) and (u,v) respectively. It should always be remembered, however, that when this loop is implemented in practice both the image and its visibility are sampled. An estimate of the true image is denoted by $k(x,y)$.

The iterative processing loops can commence in either Fourier or image space. In practice it is convenient to start in image space because it is then possible to view the image in image space after each complete iteration. In the remainder of this section the true image is assumed to be both of compact size and positive. Although it is essential to have at least one constraint in image space, the loop can proceed by enforcing any number of constraints in image space.

The initial estimate of the true image, $k_0(x,y)$, is thus chosen (in a manner discussed further in §5.1) to be both of the correct size and positive. The visibility of this initial estimate does not, however, agree with the known information in Fourier space. The next stage forces $K_i(u,v)$ to agree with the known data in Fourier space. For the phase problem this involves modifying $K_i(u,v)$ such that (where $\mathcal{P}[\ ]$ again stands for phase of)

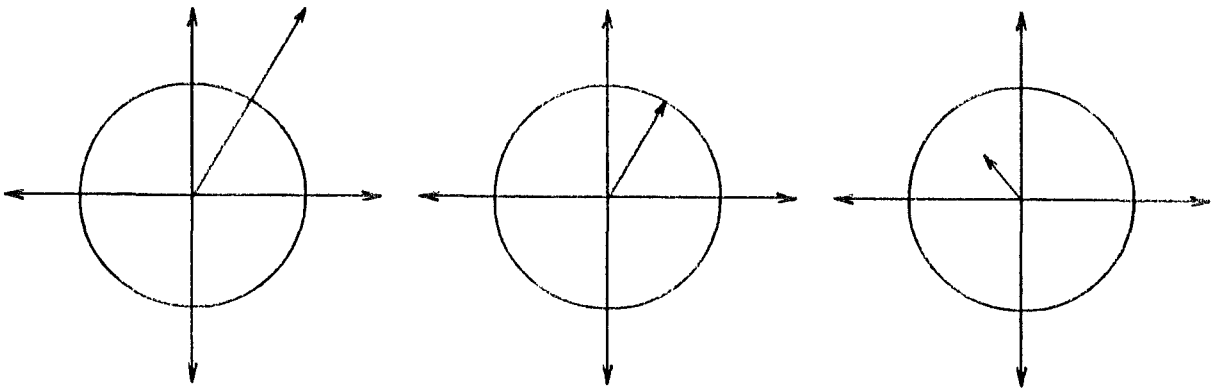
$$K'_i(u,v) = |F(u,v)|\mathcal{P}[K_i(u,v)] \tag{5.1}$$

whilst for the magnitude problem this entails

$$K_i'(u,v) = |K_i(u,v)|\mathcal{P}[F(u,v)] \tag{5.2}$$

The next stage involves returning to image space by the inverse Fourier transform. This new estimate, $k_i'(x,y)$ does not in general agree with the support constraint in image space, i.e. it has non-zero pixels outside the support or negative pixels. $k_i'(x,y)$ is then forced to comply with the image space constraints to produce the starting image for the next iteration, $k_{i+1}(x,y)$. The differences between the majority of iterative algorithms are in the application of the image space constraints.

It is also possible to view the imposition of constraints in Fourier space vectorially. Fig 5.2 shows, for the phase problem, the estimate of magnitude and phase of a point in Fourier space at the various stages of the iterative loop. In general $K_i(u',v')$ does not have the correct Fourier magnitude. The locus of possible values of $K_i(u',v')$ meeting the Fourier magnitude constraint is a circle centred on the origin. Replacing the erroneous Fourier magnitude with the known value of $|F(u',v')|$ effectively chooses from the locus of possible values defined by $|F(u',v')|$ the point closest to $K_i(u',v')$. This new estimate $K_i'(u',v')$ now meets the Fourier constraints but violates the image space constraints. Application of the image space constraints again alters $|K_{i+1}(u',v')|$, necessitating further application of the Fourier constraints, but more importantly $\mathcal{P}[K_{i+1}(u',v')]$ is (usually) closer to $\mathcal{P}[F(u',v')]$.



$K_i(u',v')$ $K_i'(u',v')$ $K_{i+1}'(u',v')$

(a) (b) (c)

Figure 5.2: Effect of application of constraints on a point in Fourier space. The locus defined by the true magnitude $|F(u',v')|$ is the circle in (a), (b) and (c). (a) The initial estimate at the start of the loop, $K_i(u',v')$. Note that the Fourier magnitude is incorrect. (b) After the application of the Fourier magnitude constraint. $\mathcal{P}[K_i'(u',v')] = \mathcal{P}[K_i(u',v')]$. (c) After the application of the image space constraints. Application of the image space constraint usually results in an updated estimate of $\mathcal{P}[F(u',v')]$ since $\mathcal{P}[K_{i+1}(u,v)] \neq \mathcal{P}[K_i(u',v')]$ Note that the Fourier magnitude of $K_{i+1}(u',v')$ is again incorrect.

For convenience the symbol Ξ is used to denote the set of points in image space where the current estimate $k(x,y)$ violates the known constraints in image space. Thus

when dealing with a positive image of compact support Ξ includes all points outside the support and those points within the support where where $k(x,y)$ is negative. The oldest and simplest method of making $k_i'(x,y)$ agree with the image space constraints is to set all pixels violating the constraints in images space to zero:

$$k_{i+1}(x,y) = k_i'(x,y) = \begin{cases} k_i'(x,y) & x \notin \Xi \\ 0 & \in \Xi \end{cases} \tag{5.3}$$

It is worth noting that this operation is non-expansive, i.e. $k_{i+1}(x,y)$ cannot be further from the true solution than $k_i'(x,y)$ (Youla and Webb 1982). The advantage of non-expansive operations is that the convergence of the iterative loop can be analysed in a simple manner and some useful results obtained (Tom et al 1981). Unfortunately, the application of constraints in this manner can lead to very slow convergence of the iterative loop.

The above discussion has mentioned convergence of an iterative loop. Ideally this convergence would be measured by the difference between the true image and the estimate. When one is presented with actual (as opposed to simulated) data, however, the true image is unavailable. In simulations, comparing the reconstruction with the true image provides a useful benchmark for assessing the other error measures discussed below. The true error or E_T is thus defined as

$$E_T = \frac{\int_{(x,y)} |f(x,y) - k_{i+1}(x,y)|^2 \, dxdy}{\int_{(x,y)} |f(x,y)|^2 dxdy} \tag{5.4}$$

For the Fourier phase problem E_T can be correlated reasonably well with a human observers' perceptions of image quality. Hence reconstructions with a specific value of E_T are usually visually similar. It is also possible to define a level of E_T at which it is no longer possible to distinguish the reconstruction from the true image on the display device used. In the magnitude problem, however, it is necessary to somehow scale the estimate appropriately, because without knowledge of the Fourier magnitude it is impossible to estimate the correct image intensity.

When the true image is unknown (as is always the case with real world data) there are a number of error measures which can be calculated. The image space error is thus the amount by which $k_i'(x,y)$ violates the constraints in image space:

$$E_I = \frac{\int_{(x,y)\in\Xi} |f(x,y) - k_{i+1}(x,y)|^2}{\int_{(x,y)} |f(x,y)|^2 dxdy} \tag{5.5}$$

Clearly this error would be zero if $k_i(x,y)$ were equal to $f(x,y)$. Furthermore, if this were the case, the image would remain unchanged by another iteration, as would be expected of the true image. Alternatively it is possible to calculate how much $K_i(u,v)$ violates the Fourier constraints. For example in the phase problem the measure

$$E_F = \frac{\int_{(u,v)} (|F(u,v)| - |K_i(u,v)|)^2}{\int_{(u,v)} |F(u,v)|^2 dxdy} \tag{5.6}$$

can be calculated.

In all processing loops it is necessary to make some form of initial estimate of the image. There have been a number of techniques proposed for making starting estimates (Won et al. 1985, Fienup and Wackerman 1986). There are some particularly poor starting estimates which should be mentioned. A symmetric starting estimate results in an entirely real Fourier transform, which if the support constraint is also symmetric, constrains the estimate to be a symmetric function. Similarly, totally asymmetric choices of initial estimate should also be avoided.

Since the convergence of the iterative loop is a strong function of the initial estimate it is essential to run comparative simulations with either the same starting image or better still a number of different starting images. A random number sequence has the advantage that several different starting points can easily be tried, in keeping with the philosophy that the final reconstruction should ideally be insensitive to changes in any assumptions necessary to iteratively process the actual data.

5.2 Estimation of the support in image space

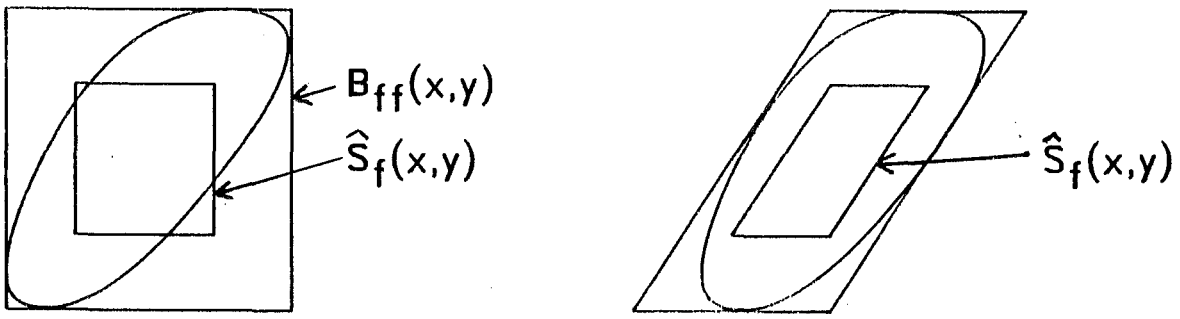
In the Fourier phase problem the only known constraint is the Fourier magnitude which is sampled at twice the Nyquist frequency of the true image (§1.3), thus enabling an estimate of $ff(x,y)$, the autocorrelation of the true image, to be formed. In order to apply a support constraint in image space it is necessary to deduce the image support from the support of its autocorrelation. It should be remembered that it is not possible to determine the true image exactly, but only its image form (§3.2). Because the image-form is invariant under translation it is only possible to estimate the size, and not the position of the image support.

When estimating $S_f(x,y)$ from $S_{ff}(x,y)$ it is only possible to rigorously obtain an upper bound on the size of a positive image (Fienup et al. 1982). It is more difficult to obtain the support of a complex or bipolar image. Reference to (1.29) helps to make this clear. If $f(x,y)$ is positive then the integrand in (1.29) is also entirely positive. Hence $ff(x,y)$ is only zero when the integrand is identically zero for all points in image space. The dimensions of a convex region bounding $S_{ff}(x,y)$ are at most twice the dimensions of a convex region bounding $S_f(x,y)$ (Bates and McDonnell 1986). Using the notation of §1.4 it is seen that

$$L_{ff}(x_k) = 2 \cdot L_f(x_k) \tag{5.7}$$

By contrast a complex or bipolar image can be imagined to oscillate such that the integrand of (1.21) cancels outside a region of space smaller than twice the dimension of the image support (Bates and McDonnell 1986, §7). In my experience, however, it appears that the techniques used to estimate the support of a positive image are also usually successful when applied to the autocorrelation of a bipolar or complex image. The consequences of incorrectly estimating the size of an image-form's support in the iterative loop of §5.1 are illustrated in §§5.3 and 5.6.

There are a number of ways of estimating the support of a positive image-form from the support of its autocorrelation (Fienup et al 1982). Ideally this constraint should be as "tight" as possible, i.e. it should be larger than but as close to the true support as possible. The simplest estimate of the support is to use an image box $B_f(x,y)$ that is half the dimensions of $B_{ff}(x,y)$ as shown in Fig 5.3a. In general any parallelogram which



(a) (b)

Figure 5.3: Simple estimation of the image support from the support of the autocorrelation. The estimated support is denoted by $\hat{S}_f(x,y)$. (a) Using half the dimensions of $B_{ff}(x,y)$ (b) Using half the dimensions of an arbitrary enclosing parallelogram.

encloses the autocorrelation can be used to form an estimate of the image support, by halving its dimensions. In some cases an appropriate choice of enclosing parallegram can lead to a "tighter" (more accurate) estimate of the image support than that obtained from using half the dimensions of $B_{ff}(x,y)$. An example of this form of support estimation is shown in Fig 5.3.

A more sophisticated approach can be applied when the autocorrelation support, $S_{ff}(x,y)$, is convex (Fienup et al 1982), i.e.

$$t(x_0,y_0)+(1-t)(x_1,y_1) \in S_f(x,y) \;\; \forall t \in [0,1]; \;\; \forall (x_0,y_0),(x_1,y_1) \in S_f(x,y) \qquad (5.8)$$

The process involves determining the intersection of the autocorrelation support $S_{ff}(x,y)$ and appropriately translated versions of $S_{ff}(x,y)$. The intersection of two supports is defined as those points which are within both supports (for a more rigorous definition refer to Fienup et al. 1982).

The first step in estimating $S_f(x,y)$ is to choose a position vector (x_1,y_1) such that $(x_1,y_1) \in S_{ff}(x,y)$. The intersection of $S_{ff}(x,y)$ and $S_{ff}(x+x_1,y+y_1)$ provides an upper bound, denoted by $I_2(x,y)$, on the size of $S_f(x,y)$. Thus when $S_f(x,y)$ is appropriately translated,

$$I_2(x,y) = (S_{ff}(x,y) \cap S_{ff}(x+x_1,y+y_1)) \;\; \supset \;\; S_f(x+x_3,y+y_3) \qquad (5.9)$$

This estimate of the image support can be further improved by, firstly, choosing a second position vector (x_2,y_2) such that $(x_2,y_2) \in (S_{ff}(x,y) \cap S_{ff}(x_1,y_2))$ and secondly forming the triple intersection, denoted by $I_3(x,y)$, of $S_{ff}(x,y)$, $S_{ff}(x+x_1,y+y_1)$ and $S_{ff}(x+x_2,y+y_2)$. Consequently

$$I_3(x,y) = (S_{ff}(x,y) \cap S_{ff}(x+x_1,y+y_1) \cap S_{ff}(x+x_2,y+y_2)) \;\; \supset \;\; S_f(x+x_3,y+y_3) \quad (5.10)$$

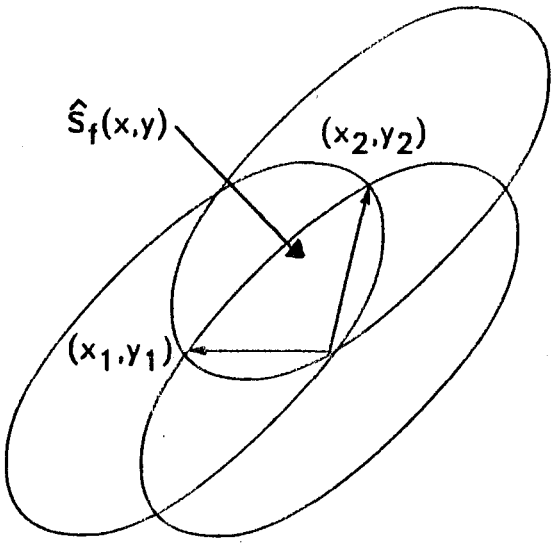


Figure 5.4: Estimation of the image support from the support of the autocorrelation using autocorrelation tri-intersection for convex sets.

A graphical interpretation of this is shown in Fig 5.4.

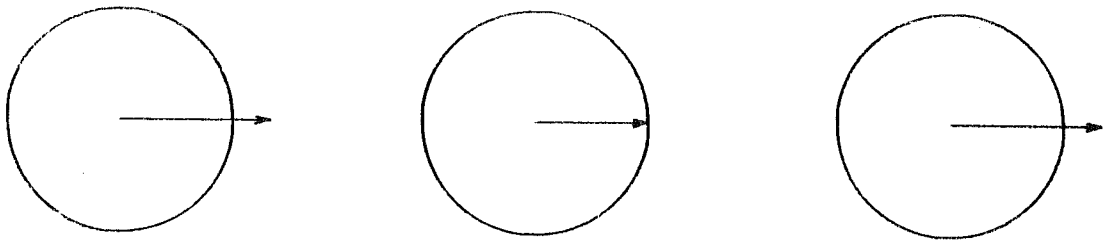
It is important to realise that all of these techniques for estimating $S_f(x,y)$ from $S_{ff}(x,y)$ usually overestimate the size of the image-form's support. In special cases, for example when the image consists of discrete points, it is possible to exactly determine the image-form support from the support of its autocorrelation (Fright 1984, Fienup et al 1982), but these cases are the exception rather than the rule.

In practice, determination of the support is made more difficult by the presence of noise. As described in §1.4 it is usual to define the edges of the support to be where the image is of smaller amplitude than the background noise. In general, this causes the autocorrelation support to be underestimated, which to some degree counteracts the overestimation mentioned in the previous paragraph. There is, however, inevitably some uncertainty in the estimate of the image support when using real world data.

In general, rather than employ a single support constraint, it is usual to try several different sizes of support. Too small a support usually results in there being no image-form compatible with the Fourier space constraints. On the other hand, small overestimates of the support usually result simply in slow convergence. By trying the iterative process with supports of different sizes it is possible to ensure the insensitivity of the reconstruction to assumptions made about the exact size of the support.

5.3 Iterative recovery of Fourier phase

Formation of an iterative loop using the Fourier constraint (5.1) and the image space constraint (5.3) results in what is called in this thesis the error reduction algorithm (Fienup 1982). This algorithm is identical to that analysed by a number of authors (Hayes 1982; Oppenheim and Lim 1981). Perhaps the most desirable aspect of this algorithm is



$K_i(u',v')$ (a) $K_i'(u',v')$ (b) $K_{i+1}'(u',v')$ (c)

Figure 5.5: Illustration of stagnation due to symmetric estimate of the true image coupled with symmetric image space constraints. Shown is the estimate of the Fourier phase at a point (u',v'). Note that $\mathcal{P}[K_i(u',v')] = \mathcal{P}[K_{i+1}(u',v')]$. (a) $K_i(u',v')$ (b) $K_i'(u',v')$ (c) $K_{i+1}(u',v')$.

that it is easy to analyse. The most serious drawback is that its convergence is usually too slow to be employed in practice.

The analysis of the error reduction algorithm (Fienup 1982) shows it is equivalent to a steepest descent search on E_I (which in the case of the error reduction algorithm can be related through Parseval's theorem, §1.6, to E_F). Furthermore, it can be shown that the error reduction algorithm always results in a reduction in E_I. Use of a steepest descent search does not, however, guarantee convergence to a global minimum. As an example consider the problem of recovering a non-symmetric positive image-form from its Fourier magnitude. As noted previously, if both the initial estimate $k_0(x,y)$ and the support constraint $S_f(x,y)$ are symmetric, neither application of the image nor the Fourier constraints can cause the estimate to become asymmetric. Thus full convergence can never occur in such a case.

Iterative algorithms can fail in two ways. The first is divergence, when the reconstruction actually becomes further from the true image as more iterations are employed. The second, and more common failure, is the onset of stagnation. Stagnation occurs when the reconstruction loop makes very little progress towards the true image despite large numbers of iterations. In some cases, progress is eventually made, but there are situations where the true image can not be obtained even with an indefinite number of iterations.

The stagnation that occurs when $k_i(x,y)$ and $S_f(x,y)$ are symmetric and the true image is symmetric is shown in Fig 5.5. Since the Fourier transform of a conjugate symmetric image is always real, application of a symmetric support constraint merely causes an alteration in the magnitude of the Fourier transform, Fig 5.5a. When the Fourier magnitude constraint is applied this merely reverses the change made by the image space constraint, Fig 5.5b. A complete passage through the iterative loop causes no progress to the true phase as illustrated in Fig 5.5c (cf Sanz et al. 1984).

This problem of stagnation (or locking) has been also noted in related phase retrieval problems. The Gerchberg-Saxton algorithm (Gerchberg 1986, Gardenier and

Bates 1987) is employed to recover complex images when both the Fourier and image magnitudes are known. The Gerchberg-Saxton technique is essentially the same as the error reduction algorithm, except for application of the image space constraints. Instead of enforcing a support constraint, the image magnitude is set to the known magnitude in image space. The added information of the image space magnitude usually results in vastly superior convergence when compared with the normal error reduction algorithm. Even though it can be shown that the errors in both Fourier and image space must decline, stagnation still occurs in this algorithm. Thus even when the image magnitude is known, it is advisable to invoke more sophisticated algorithms (a topic currently under investigation by Gardenier at the University of Canterbury).

It is very important to note that a monotonically decreasing error curve does not guarantee convergence even if a unique minimum does exist. For instance the sequence defined by

$$0.5 + \frac{1}{2^n} \quad n = 0, 1, 2 \ldots \tag{5.11}$$

also has a monotonically decreasing nature but never falls below 0.5.

The above difficulty can be averted in the Fourier phase problem by ensuring $k_0(x, y)$ is neither totally symmetric, nor totally asymmetric. Even when this is done, however, the convergence of steepest descent algorithms is often very slow (Fletcher 1983). The stagnation of a steepest descent usually occurs when the minimisation problem is ill-conditioned, which seems almost certainly to be the case in most phase retrieval examples reported in the literature (it is certainly true for the great majority of examples I have tried).

When describing the stagnation of a steepest descent minimisation technique it is useful to introduce the concept of the step made in an iteration. During one iteration the change in the reconstruction (or step taken) is given by

$$\Delta k_i = k_i'(x, y) - k_{i-1}'(x, y) \tag{5.12}$$

Fig 5.6 illustrates how a steepest descent search can become stagnated when E_I is far removed from the global minimum (and consequently the reconstruction is also considerably different from the true image). Although the step taken is quite large, little overall progress is made towards the true solution, because the direction of the minimum and the direction of steepest descent are approximately orthogonal.

There are a number of ways of overcoming the problems inherent in using a steepest descent minimisation technique. The first is to describe the problem in such a manner that it is no longer ill-conditioned, This has led Newsam and Barakat (1986) to propose an algorithm based on representing the image by continuous prolate spheroidal wave functions. To the best of my knowledge this technique has not produced any concrete examples of two-dimensional phase recovery, possibly due to difficulties associated with calculating continuous prolate spheroidal wavefunctions. An alternative suggestion based on using discrete prolate spheroidal wave functions is presented in chapter 7.

The traditional method of dealing with ill-conditioned minimisation problems has been to adopt a more sophisticated minimisation technique. Second order or quasi-Newton methods (Fletcher 1983) have proven effective when dealing with ill-conditioned minimisation problems. The virtue of these algorithms is that they model the error surface with a second order function and so can take curvature information into account. The major drawback of second order methods is the severe computational burden that is

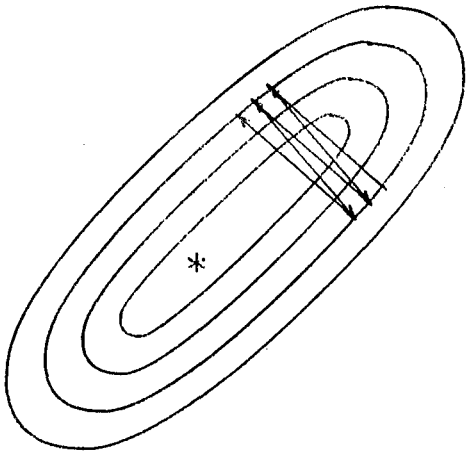


Figure 5.6: Schematic illustration of how stagnation can occur in a steepest descent search. Note that the direction of steepest descent does not point to the global minimum which is marked with an asterisk.

inevitably associated with them, mainly because they require matrix calculations. The number of elements in these matrices is equal to the square of the number of pixels in the image, which leads to unacceptably large amounts of computation (Fienup 1982).

A simple possibility is to employ a line search, rather than just taking a fixed step in the direction of steepest descent. By defining

$$k_i''(x,y) = k_i'(x,y) + \Lambda(k_i'(x,y) - k_{i-1}'(x,y)) \tag{5.13}$$

it is possible to optimise the value of Λ to give the greatest reduction in the image space error (in traditional error reduction the value of Λ is fixed at 1.0). The starting estimate for the next iteration is then constructed by applying the image space constraints to $k_{i+1}''(x,y)$ rather than $k_{i+1}'(x,y)$. Λ is optimised because it is possible, as shown in Fig 5.6, that taking a fixed step may overshoot (or undershoot) the minimum functional value along the path of steepest descent. Thus although the error reduction guarantees a reduction in E_I, it is possible that a larger reduction could be obtained if a smaller (larger) value of Λ was used. Further acceleration of convergence can be achieved by employing a conjugate gradient search (Sasaki and Yamagami 1986, Fletcher 1983, Fienup 1982). Conjugate gradient techniques search not only in the direction of steepest descent but also in a number of orthogonal directions (thus overcoming the problem illustrated in Fig 5.6).

The problem with the above techniques is, although they are capable of finding a minimum quickly, there is no guarantee that this is the global minimum. All variations of descent search algorithms, which include quasi-Newton and conjugate gradients, are ineffective in the presence of multiple local minima. This is because, when the reconstruction corresponds to a local minimum in E_I, there is no guarantee that the reconstruction is in fact close to the true image. In the Fourier phase problem it appears that interaction

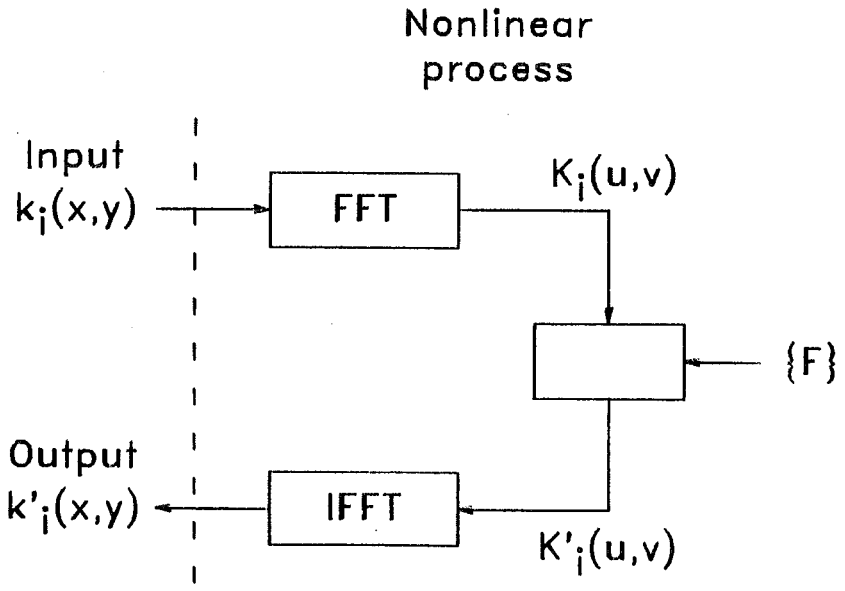


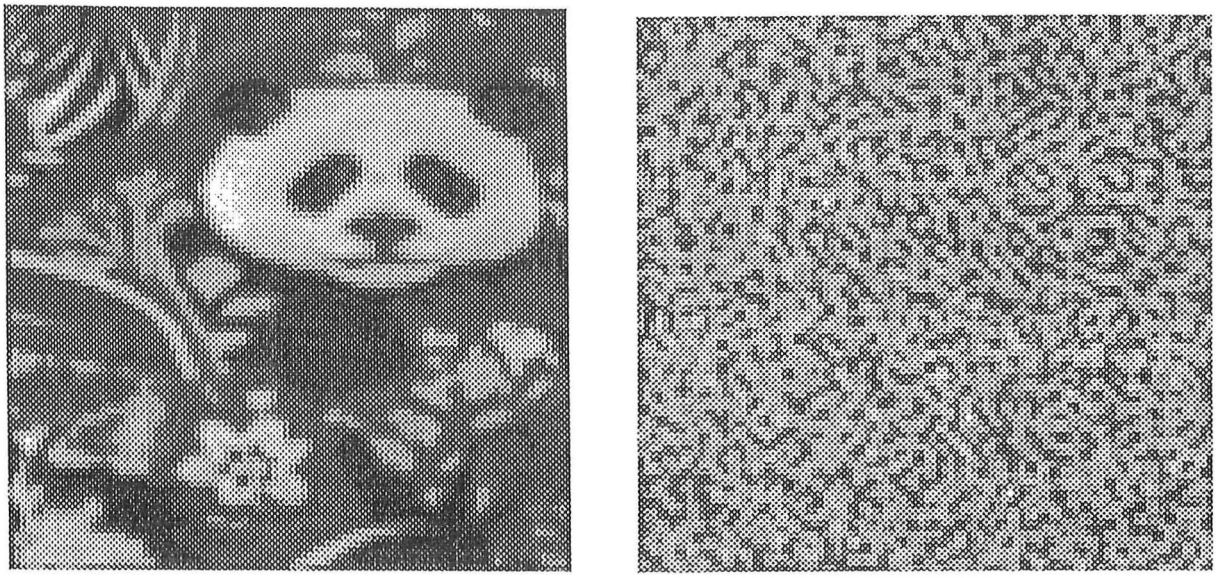
Figure 5.7: Block diagram showing the input-output view of applying the Fourier magnitude constraint in the phase problem.

between the image and Fourier space constraints can introduce local minima, a point illustrated in Figs 5.5.

The most practical modification of the error-reduction algorithm is Fienup's hybrid input-output algorithm. The reasoning behind this algorithm is somewhat heuristic, but it has proved very effective in computational examples and in practice (Lane 1987, Fright 1984, Fienup 1982, Feldkamp and Fienup 1980). The derivation of the hybrid input-output algorithm involves viewing the application of the magnitude constraint as a non-linear process, Fig 5.7. It is informative to begin discussion of the hybrid input-output algorithm with an examination of the pure input-output algorithm (Fienup 1982). The basis of the latter algorithm is that a small change in the input, $k_i(x,y)$, usually results in a similar change in the output $k_i'(x,y)$. This reasoning suggests the following method of applying the image constraints

$$k_{i+1}(x,y) = \begin{cases} k_i(x,y) & (x,y) \notin \Xi \\ k_i(x,y) - \beta k_i'(x,y) & (x,y) \in \Xi \end{cases} \tag{5.14}$$

where β is feedback parameter. The input-output algorithm, essentially tries to synthesise an input which after being passed through the nonlinear process meets the image space constraints. The input $k_i(x,y)$ may not meet the constraints in either image or Fourier space, since it is merely a function that when passed through the non-linear process of Fig 5.7 produces an output, $k_i'(x,y)$, which does meet these constraints. Consequently, only E_I can be used to measure algorithm convergence, because E_F can still be large as complete convergence only requires that $\mathcal{P}[K_i(u,v)] = \mathcal{P}[F(u,v)]$. It is worth noting that a common mistake in implementing input-output algorithms is to implement the image space constraints in a manner akin to (5.13), rather than (5.14).



(a) (b)

Figure 5.8: Positive 64 x 64 pixel images quantised to 32 grey levels, ranging from 0 (black) to a normalised value of 1 (white). (a) true image (b) pseudo random starting estimate.

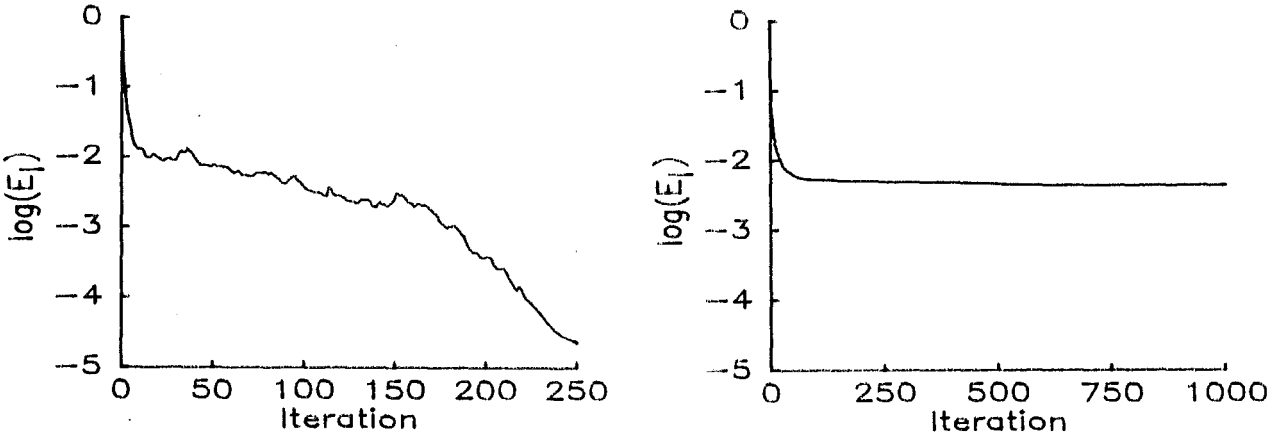
The most effective method combines the input-output and error-reduction methods and chooses a new estimate of $k_{i+1}(x,y)$ by

$$k_{i+1}(x,y) = \begin{cases} k_i'(x,y) & (x,y) \notin \Xi \\ k_i(x,y) - \beta k_i'(x,y) & (x,y) \in \Xi \end{cases} \tag{5.15}$$

which defines the hybrid input-output method (Fienup 1978). Again, for the reasons given in the previous paragraph, convergence of this algorithm can only be measured using E_I.

Although the hybrid input-output algorithm appears the best current method of phase retrieval, analysis of its convergence is by no means simple. Unlike error-reduction there is no simple analogue with steepest descent minimisation, and consequently no guarantee that E_I decreases at each iteration. The algorithm does have the advantage that stagnation is impossible. For example, when recovering a positive image, if any pixel in the output image remains negative, the feedback eventually increases the level of the corresponding pixel in the input image until the output image pixel becomes positive. This may, however, cause the overall E_I to rise by causing other pixels in the image to be more in error.

As an example of the difference between error-reduction and hybrid input-output consider the positive 64 x 64 pixel image shown in Fig 5.8a. The data are the Fourier magnitude (oversampled by a factor of two) and the a priori knowledge that the image is positive. The image size can be estimated from the data (cf §5.2) so that a support constraint can be set up in image space. The starting estimate used for both methods is the pseudo random image shown in Fig 5.8b. The contrast in performance between hybrid input-output and error reduction is dramatically illustrated in Figs 5.9, 5.10 and 5.11. Fig 5.9a shows the decline in E_I (henceforth called the error curve) for the hybrid input



(a) (b)

Figure 5.9: Plots of E_I for increasing iteration for reconstructions of the image shown in Fig 5.8. (a) Hybrid input-output, $\beta = 0.5$. (b) Error reduction.

output algorithm with $\beta = 0.3$, whilst Fig 5.9b shows the error curve for error reduction. Reconstructions for increasing numbers of iterations are shown in Figs 5.10a-d and 5.11a-d. The superiority of the hybrid input-output over the error reduction algorithm is readily apparent.

For the error reduction algorithm, the error curve is monotonically decreasing with a sharp initial descent followed by stagnation. This behaviour has been noted by a number of authors (Hayes 1982, Fienup 1982, Fiddy et al 1983) and means that error reduction is essentially useless except for contrived examples. It is important to realise that error reduction is a two stage algorithm. Application of the image constraints causes a change in the image given by

$$q_i(x,y) = k'_{i+1}(x,y) - k'_i(x,y) \tag{5.16}$$

Now consider a point (u', v') in image space. In order for $\mathcal{P}[K'_{i+1}(u,v)]$ to be different from $\mathcal{P}[K'_i(u,v)]$ it is essential for $\mathcal{P}[Q_{i+1}(u,v)]$ not to equal (or differ by π from) $\mathcal{P}[K_{i+1}(u,v)]$. If this is not so, application of the Fourier magnitude constraint reverses the effect of the image space constraints in a manner akin to that illustrated by Fig 5.5a. Fig 5.12 shows

$$\mathcal{P}\left[\frac{Q_i(u,v)}{K_i(u,v)}\right] \tag{5.17}$$

for the reconstruction of Fig 5.8a, at the iteration corresponding to the image in Fig 5.11d. Fig 5.12 shows that at stagnation there are large areas in Fourier space where the phase of the changes made by application of the image space constraints and those made by application of the Fourier space constraints are the same, or differ by π. Hence enforcement of the Fourier and image space constraints produces nearly equal and opposite changes during a single iteration and little progress is made towards the true image. This

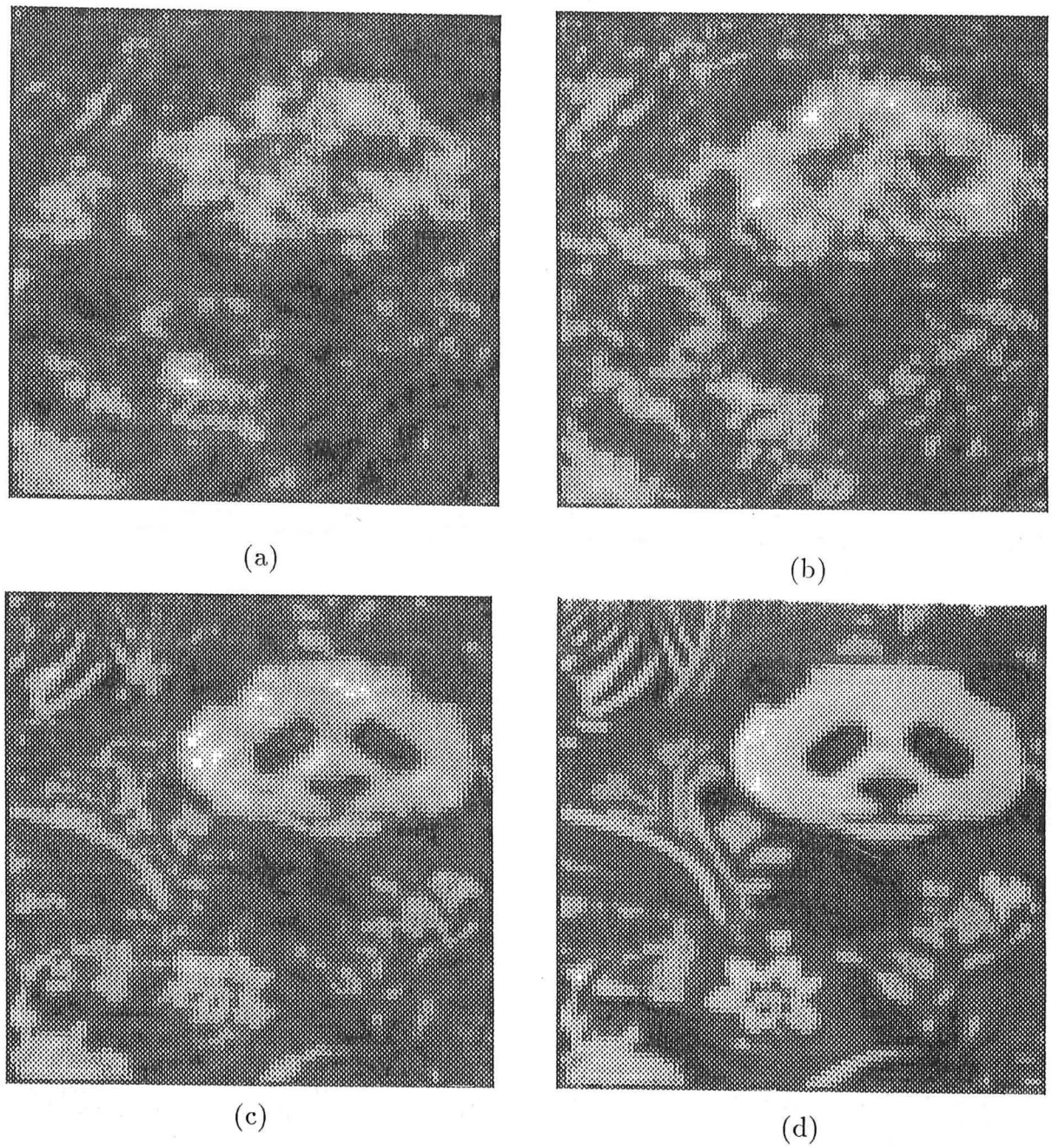


Figure 5.10: Reconstructions of Fig 5.8, using hybrid input-output, quantised as in Fig 5.8. (a) 25 iterations. (b) 50 iterations. (c) 100 iterations. (d) 250 iterations.

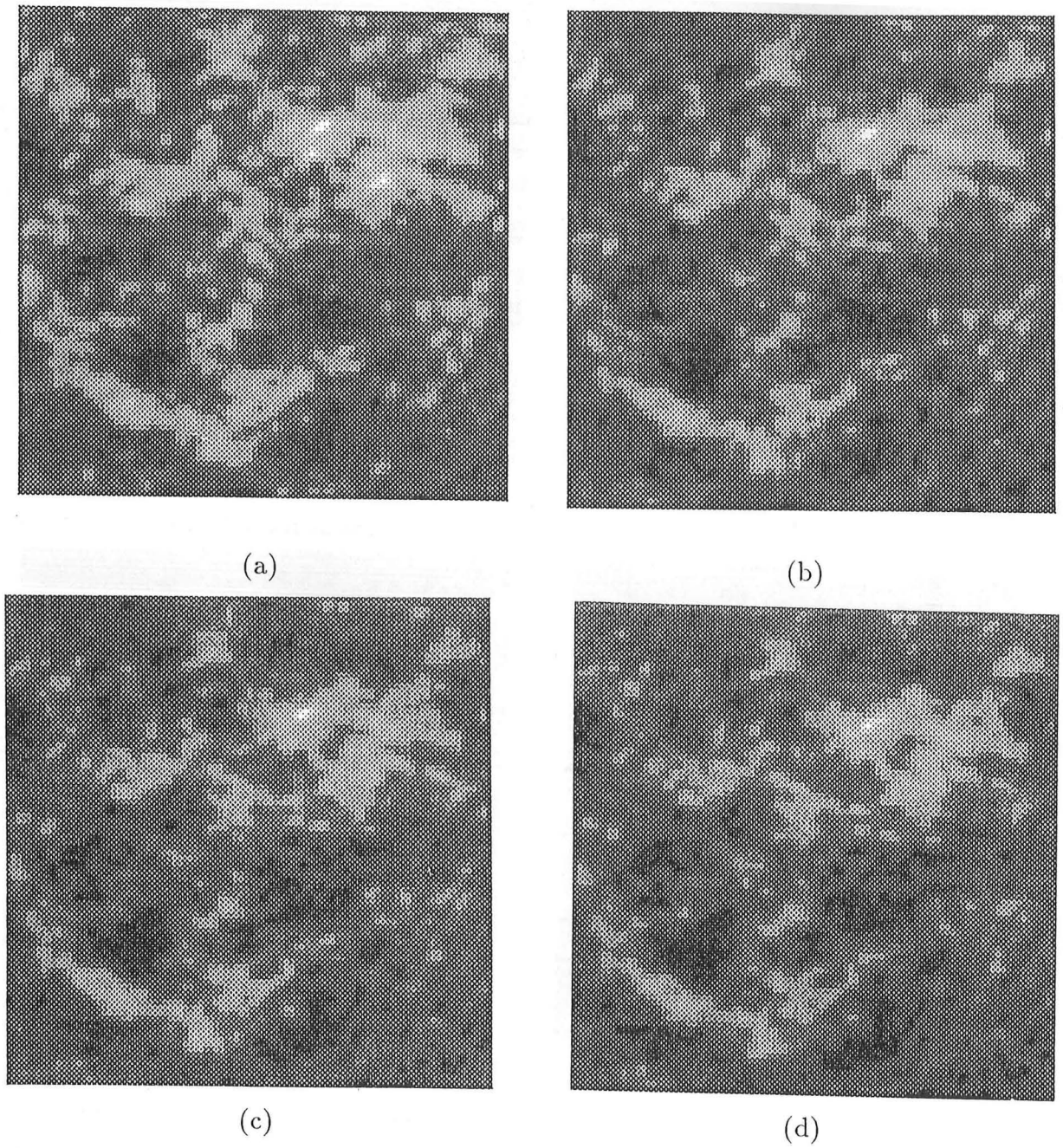


Figure 5.11: Reconstructions of Fig 5.8, using error reduction, quantised as in Fig 5.8. (a) 25 iterations. (b) 50 iterations. (c) 100 iterations. (d) 250 iterations.

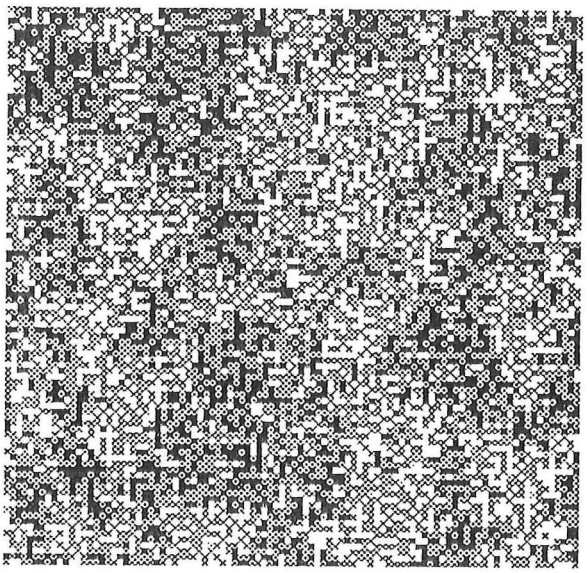
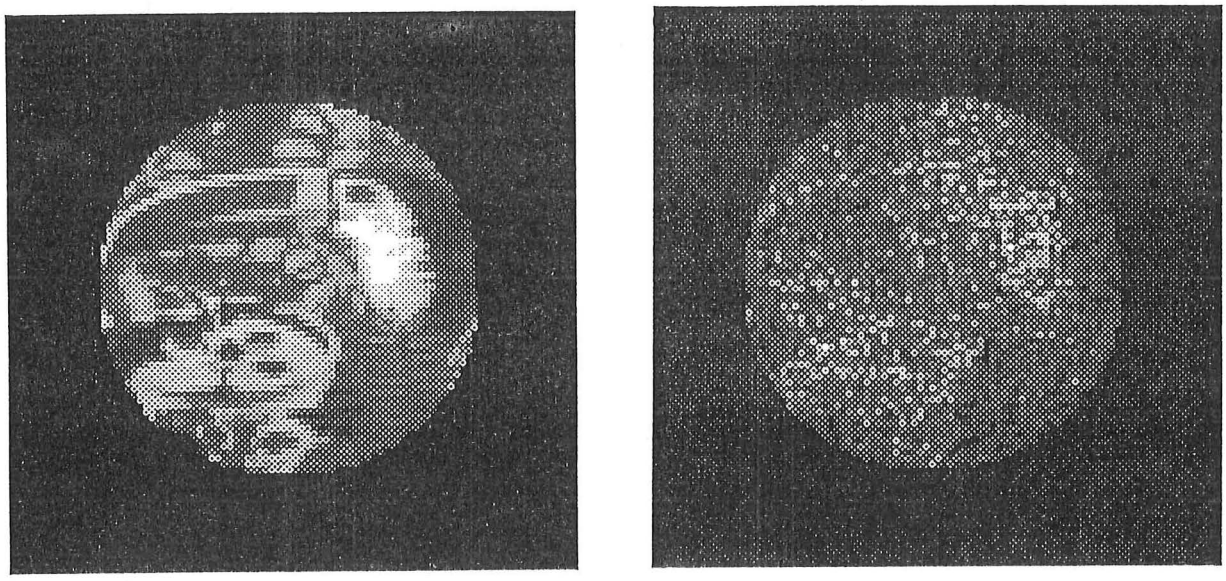


Figure 5.12: The phase of the ratio of $Q_i(u,v)/K_i(u,v)$ for the image shown in Fig 5.11d. Quantised to 32 grey levels, ranging from -π (black) to π (white), zero represented by the middle shade of grey.



(a) (b)

Figure 5.13: Magnitude of complex images quantised as in Fig 5.8. (a) true image. (b) reconstruction after 1500 iterations using a mixed approach of hybrid input-output and error reduction.

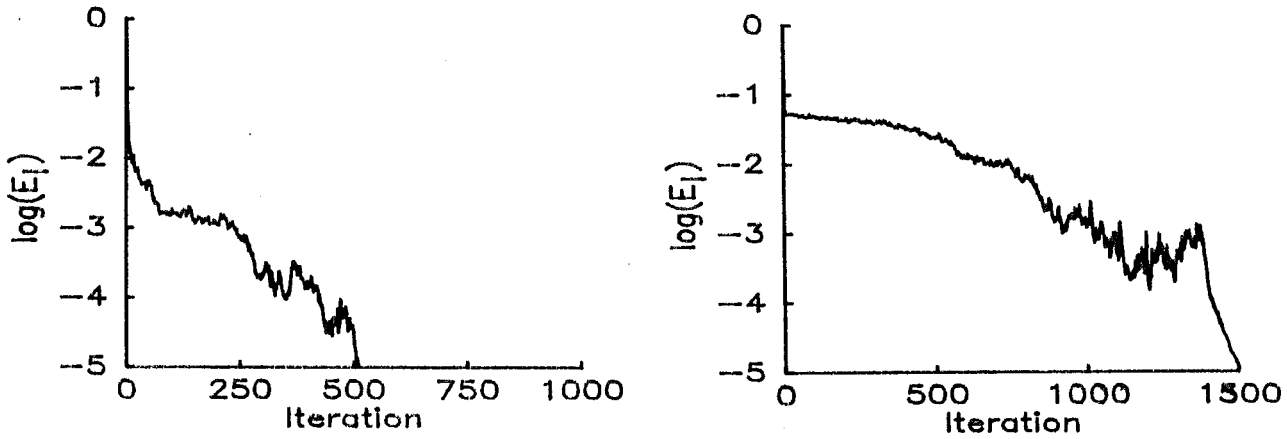
effectively produces a local minimum or stationary point. Thus, although there remains a large E_I and the reconstruction is far from the true image, there is little difference between $k_i(x,y)$ and $k_{i+1}(x,y)$.

Also worthy of note is the difference in the visual quality of reconstructions generated by the hybrid input-output and error reduction algorithms for a given value of E_I (Lane 1987, Fienup 1982). It is clear that for a given E_I, the reconstruction by hybrid input-output is much closer to the true image than the error reduction reconstruction (for example, refer to Figs 5.10c and 5.11c). In fact, when using hybrid input-output, a sharp drop in E_I can always be obtained by inserting some error-reduction iterations, but usually with no visual improvement of the image or substantive decline in E_T. In the interests of keeping E_I as low as possible it is common to employ a mixed approach of hybrid input-output and error reduction.

The decline in the error introduced by error reduction is somewhat illusory, however, since error reduction soon stagnates and it is necessary to return to hybrid input-output if further substantive progress is to be made. Unfortunately, the application of hybrid input-output nearly always results in an increase of E_I to about the level before the application of error reduction (Lane 1987), implying that little has been achieved by the reversion to error reduction. As an example of this consider the image magnitude shown in Fig 5.13a which is of circular support. Two separate complex images were generated by combining firstly a constant and then a pseudo random phase (varying between $-\pi$ and π) with the image magnitude shown in Fig 5.13a. The two images generated thus provide two extreme phase distributions in image space. The latter image (whose phase varies pseudo-randomly from pixel to pixel) can be expected to be an especially critical test of Fourier phase retrieval because of the large amount of information coded in the Fourier phase (Munson and Sanz 1986). It is also worth noting that images having smooth boundaries are, in general, more difficult to reconstruct than those having sharp corners (Lane 1987, Fienup 1987).

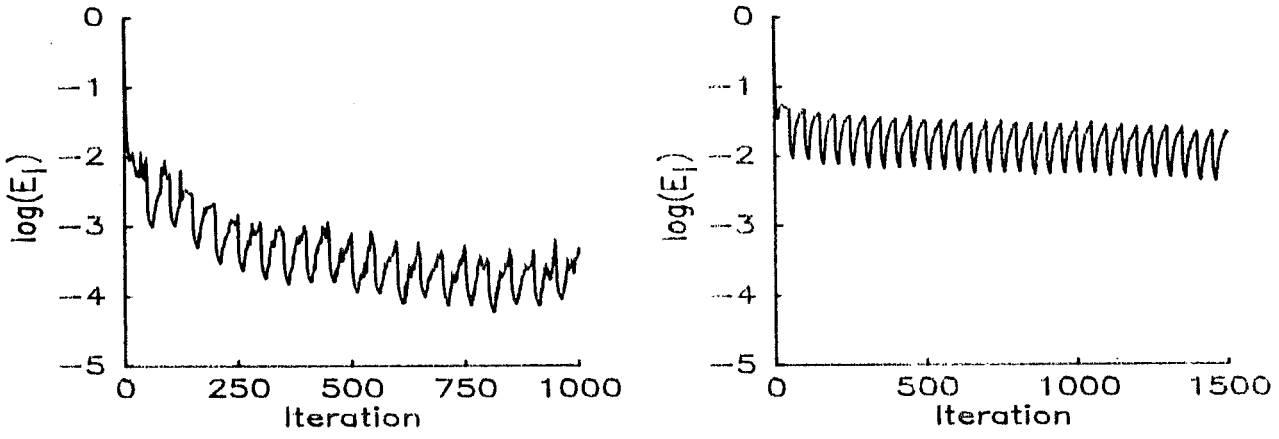
The mixed approach used to recover these two images consisted of employing cycles consisting of 10 error reduction followed by 40 cycles of hybrid input-output with a beta of 0.5. This mixed approach is contrasted with pure hybrid input-output with a beta of 0.5 in Figs 5.14 and 5.15. In both cases the same starting image was used and the only constraint employed was the support of the image. The convergence of the mixed approach is characterised by a sharp reduction in E_I whenever error reduction is applied, but with little overall progress between cycles. Thus, although both approaches successfully recovered the image form for the constant phase image within 1000 iterations, the mixed approach was significantly slower (Fig 5.14). The image with random phase required, as expected, significantly more iterations before the hybrid only approach successfully recovered the image form. Fig 5.15b shows the reconstruction using the mixed approach for the random phase image after 1500 iterations. It is apparent that significantly more cycles of the mixed approach are required before the image-form is recognisable.

In my opinion, it is apparent that the nature of hybrid input-output search is more global than that of error reduction. The error reduction algorithm appears to find a local optimum but is unable to make large changes to the reconstruction after the application of a relatively small number of iterations. Consequently, although more sophisticated minimisation techniques such as conjugate gradient and quasi-Newton methods may provide faster convergence, they are also likely to suffer from convergence to a local (as opposed to global) minimum. It is therefore important to compare E_I error curves with a degree



(a) (b)

Figure 5.14: E_I for constant phase image, magnitude as in Fig 5.13a. (a) hybrid input-output only ($\beta = 0.5$) (b) mixed hybrid input-output ($\beta = 0.5$) and error reduction.



(a) (b)

Figure 5.15: E_I for random phase image, magnitude as in Fig 5.13a. (a) hybrid input-output only ($\beta = 0.5$) (b) mixed hybrid input-output ($\beta = 0.5$) and error reduction.

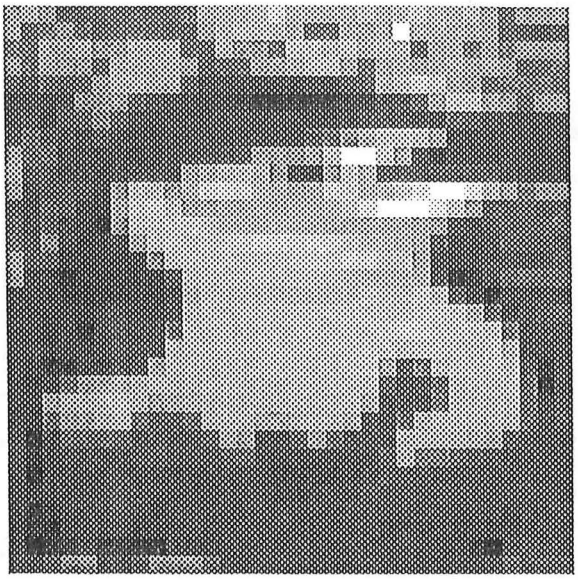


Figure 5.16: 32 x 32 pixel positive image quantised as for Fig 5.8.

of caution, because approaches employing error reduction may exhibit apparently better convergence, without necessarily getting closer to the true image.

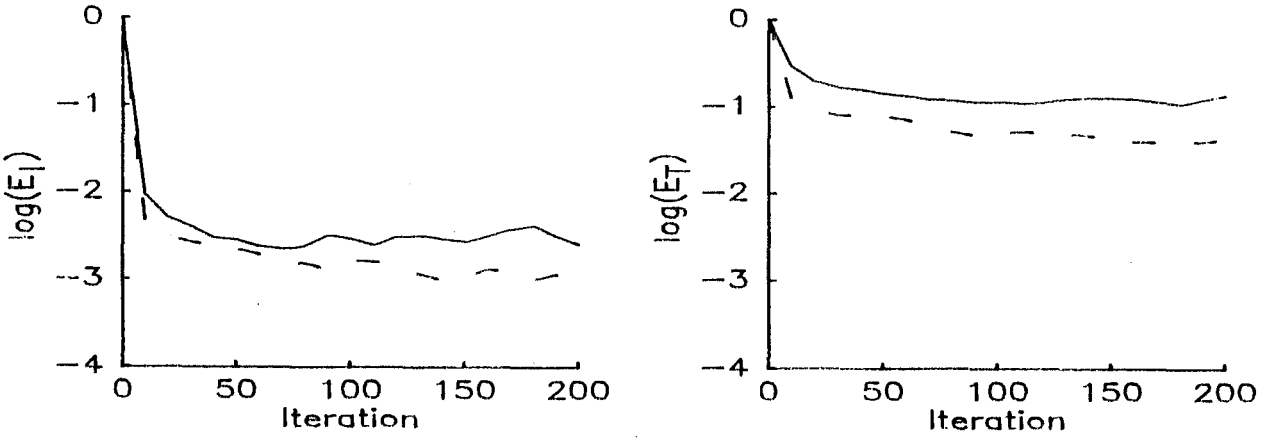
5.4 Effect of β on convergence

The rate of convergence of any of the iterative algorithms depends on a large number of factors. Categorising the behaviour of the iterative loop for all constraints, feedback parameters and initial starting estimates is a very difficult task. This section examines the effects on convergence of β, and to a lesser extent the initial starting estimate, on convergence.

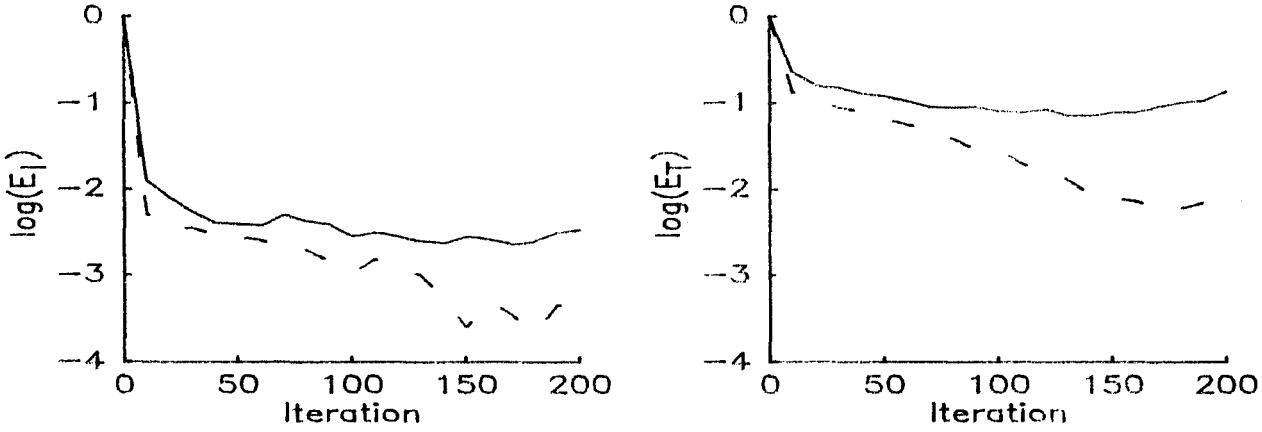
The test image employed is the positive 32 x 32 pixel image shown in Fig 5.16. The image is characterised by sharp detail which is very distinctive to a human observer. The level of E_T at which the reconstruction is deemed similar to the true image is thus much lower than the E_T required for the diffuse images reconstructed in §5.5. The image was reconstructed using hybrid input-output for three different values of β. Fig 5.17 shows the error curves for the different values of β. Five different starting images were used and the maximum and minimum values of both E_I and E_T are plotted.

The fastest convergence occurred for $\beta = 0.5$ with the slowest for $\beta = 0.1$. The lower values of β did however provide a steadier decline in the error curves. A visually acceptable reconstruction was obtained when E_T reached approximately 0.03 (or $\log(E_T)$ = -1.5) or E_I reached 0.001. There is, generally, a good correlation between the best E_I and the best E_T runs. For the runs with $\beta = 0.3$ and 0.5 the best reconstructions were all visually indistinguishable from the true image. With the $\beta = 0.3$ and $\beta = 0.5$ runs, 4 out of 5 images closely resembled the true image whilst the fifth suffered from contamination by the mirror image, a form of stagnation discussed in detail in §5.7. The $\beta = 0.1$ runs all appeared to require more iterations before convergence to a visually acceptable reconstruction occurs.

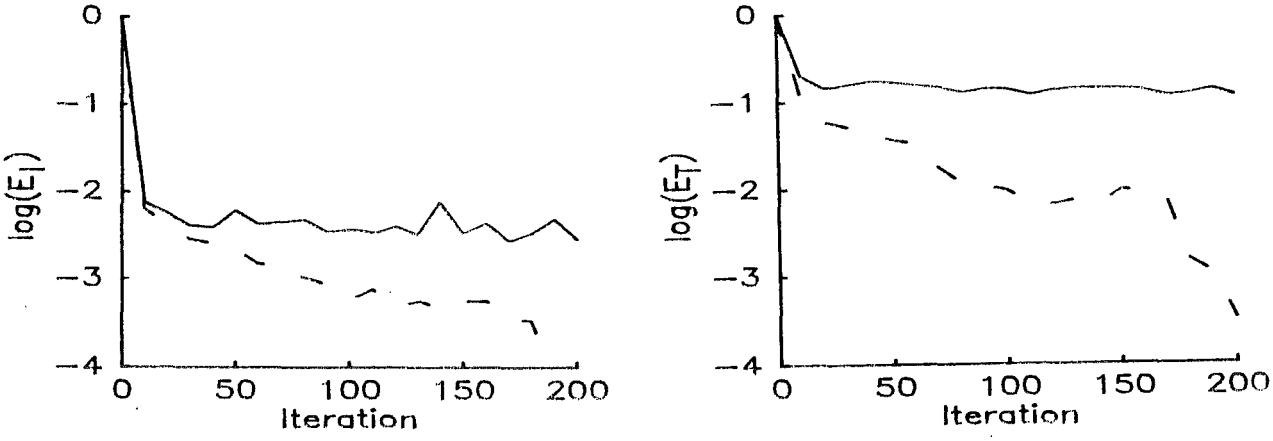
In the preceding examples the number of iterations has been curtailed for a num-



(a)



(b)



(c)

Figure 5.17: Error curves (E_I and E_T) for hybrid input-output using different values of β (and five different starting images). Dashed line shows the best reconstruction and the solid line the worst reonstruction (a) $\beta = 0.1$ (b) $\beta = 0.3$ (c) $\beta = 0.5$

ber of reasons. Some starting estimates result in stagnation which can only be overcome by large amounts of iteration. In practice, it is usually more appropriate, and is computationally less expensive, to choose a different initial starting point rather than persist with large numbers of iterations.

As an example, recovery of a 64 x 64 pixel image requires two 128 x 128 FFT's per iteration. Thus a typical positive 64 x 64 pixel image requiring approximately 200 iterations would require about 1 hour on a VAX-11 750 minicomputer. When the algorithm exhibits especially severe stagnation as many as 2000 iterations may be required before the image-form is recovered. In the absence of special purpose hardware (such as described by Fienup 1982) it would be enormously wasteful to use sufficient iterations to ensure that convergence is achieved from all starting distributions used.

It is also important to know the reasons for stagnation when dealing with noisy data. Because, in the presence of noise, there is no longer an exact solution to the Fourier phase problem, there is consequently a residual E_I. It is important to be able to distinguish between when the algorithm has reached the limits imposed by noise and when it has simply stagnated. The problems of stagnation are dealt with in detail in §5.7.

5.5 Effect of choice of support on convergence

In order to analyse the effects of different estimates of the support size on the performance of the Fienup iterations, images consisting of a number of gaussians were generated. These images are particularly difficult to reconstruct since they do not possess definite extents. Consequently, choice of a support of finite size means, even in the absence of noise, that the reconstruction is only approximate. The first quantifiable results (for "gaussian" images) involved the recovery only of images which were known to be positive (Tan and Bates 1985).

Three images were generated by assigning different phases to gaussians of fixed amplitude. The positive image was thus formed by making the phase of each gaussian zero (Fig 5.18a). For the bipolar image some gaussians were made negative (Fig 5.18b), whilst for the complex image the phases of the constituent gaussians were chosen pseudo-randomly between $-\pi$ and π, as shown in Figs 5.18c and d. Figs 5.18e,f and g show respectively the autocorrelation of Fig 5.18a, its Fourier magnitude and its Fourier phase.

The fastest reconstructions were obtained for the positive image-form. Fig 5.19 shows the best, and Fig 5.20 the worst, reconstructions from 5 different initial starting phases, whilst the error curves for the different supports are shown in Fig 5.21. In general the best reconstructions correlate well with the best E_I curves. Thus the best reconstruction, determined by a visual comparison with the true image, is usually the one with the lowest value of E_I.

The reconstructions from the 24 x 24 pixel support are clearly unsatisfactory because the image-form is too large to fit inside the estimated support. Hence even the true image generates a significant value of E_I. The best reconstructions resulted from using the 32 x 32 support, although most reconstructions using 28 x 28 and 36 x 36 pixel supports were acceptable. In general, one of the reconstructions was significantly worse than the other four and was identifiable by its high E_I.

Of more interest are the reconstructions of the bipolar image, since successful reconstruction of bipolar gaussian images has not previously been reported (cf. Tan and

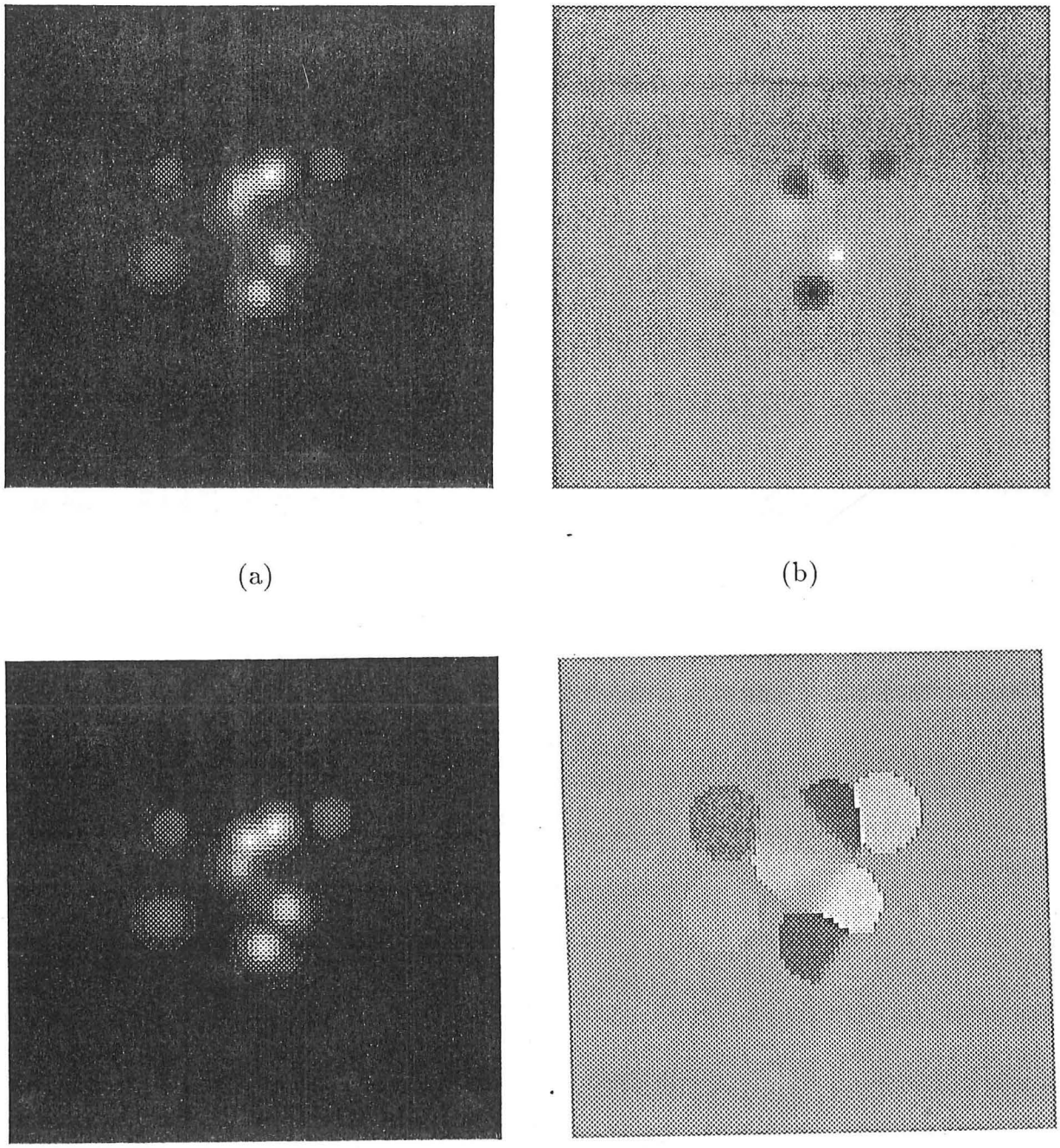
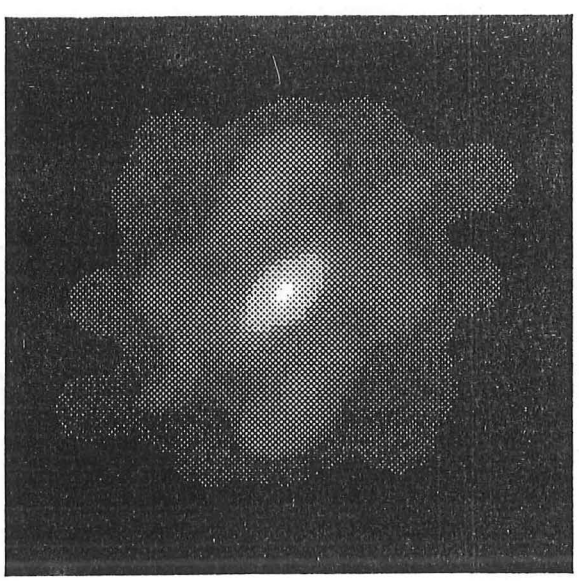
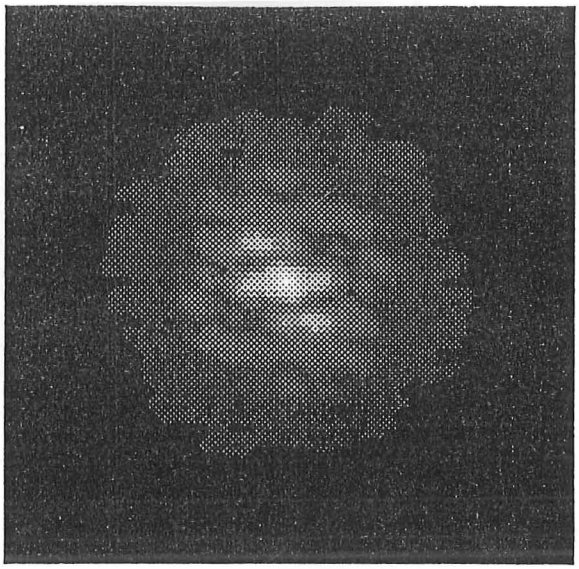


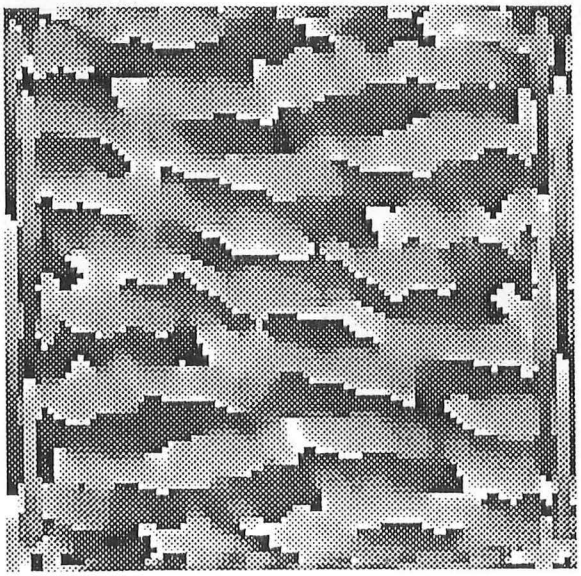
Figure 5.18: Positive, bipolar and complex images (a) Positive image quantised as for Fig 5.8. (b) Bipolar image quantised to 32 grey levels yarying from most negative (black) to most positive (white). Zero is represented by the dominant shade of (middle) grey. (c) Complex image magnitude quantised as for Fig 5.8. (d) Complex image phase quantised as for Fig 5.12.



(e)



(f)



(g)

Figure 5.18: (continued) (e) Autocorrelation of Fig 5.18a quantised as for Fig 5.8. (f) Fourier modulus of Fig 5.18a quantised as for Fig 5.8. (g) Fourier phase of Fig 5.18a quantised as for Fig 5.12.

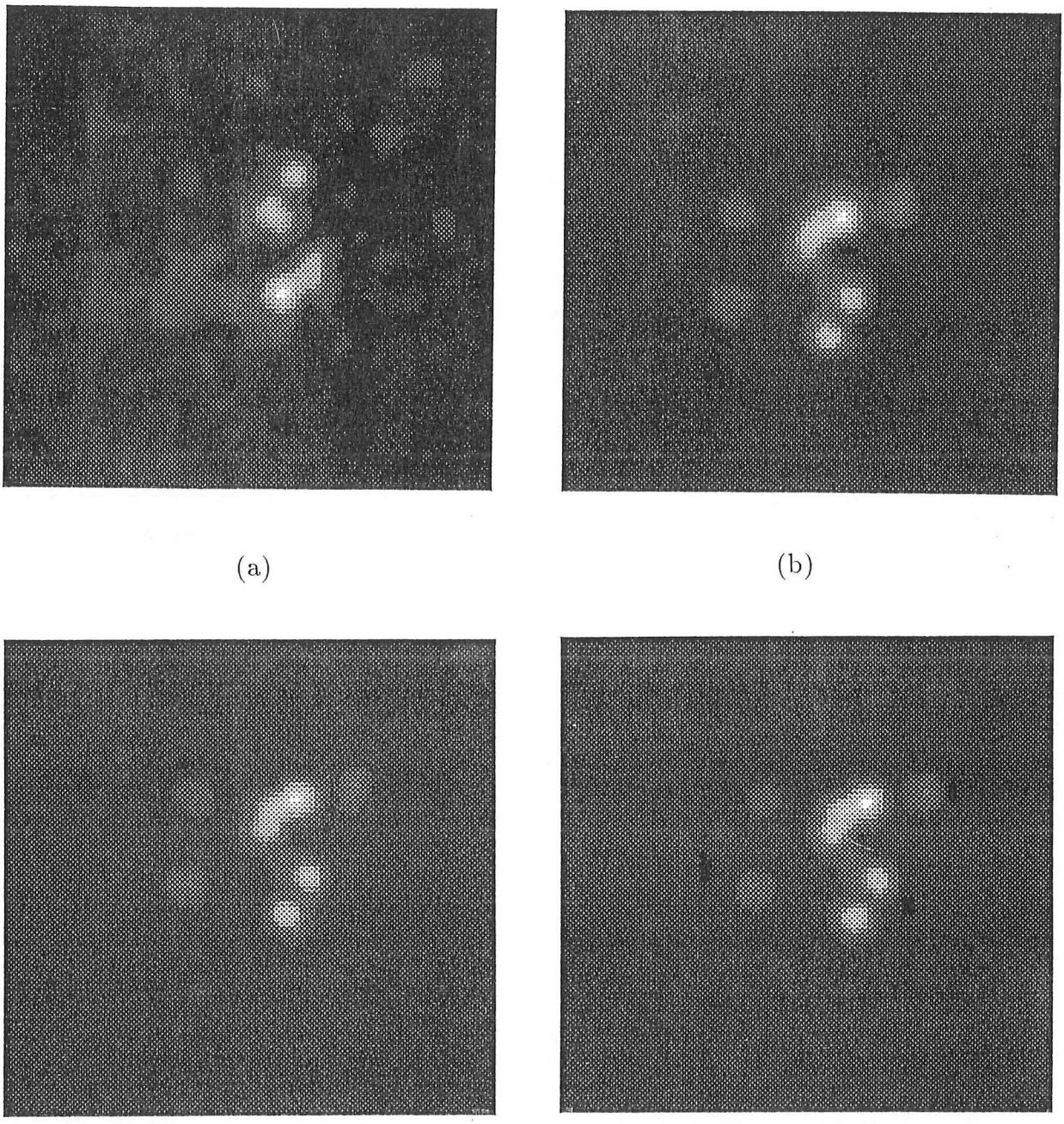


Figure 5.19: Best reconstructions (5 different starting images) of the positive image shown in Fig 5.18a, quantised as in Fig 5.8a. (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support.

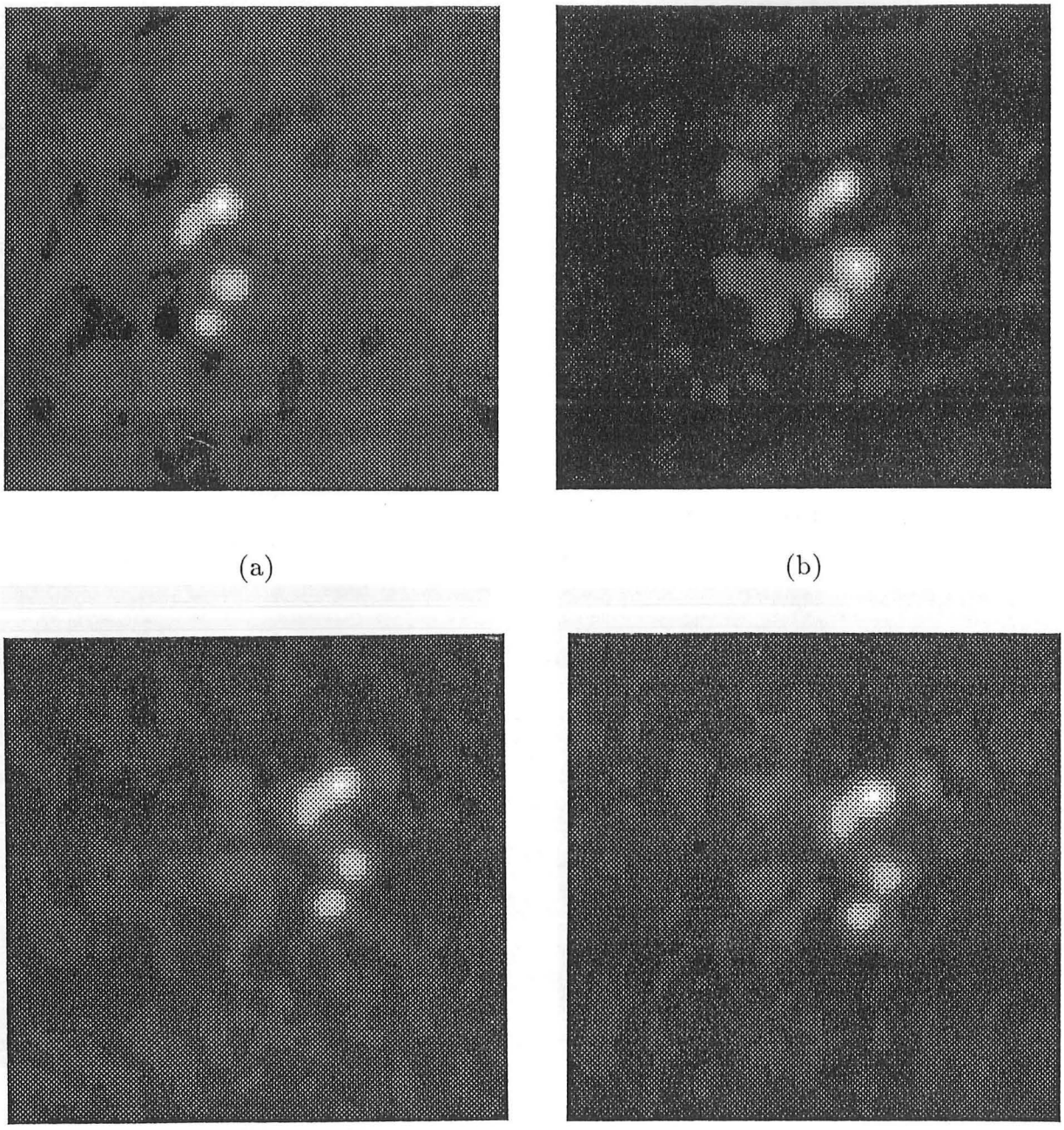


Figure 5.20: Worst reconstructions (5 different starting images) of the positive image shown in Fig 5.18a, quantised as in Fig 5.8a. (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support

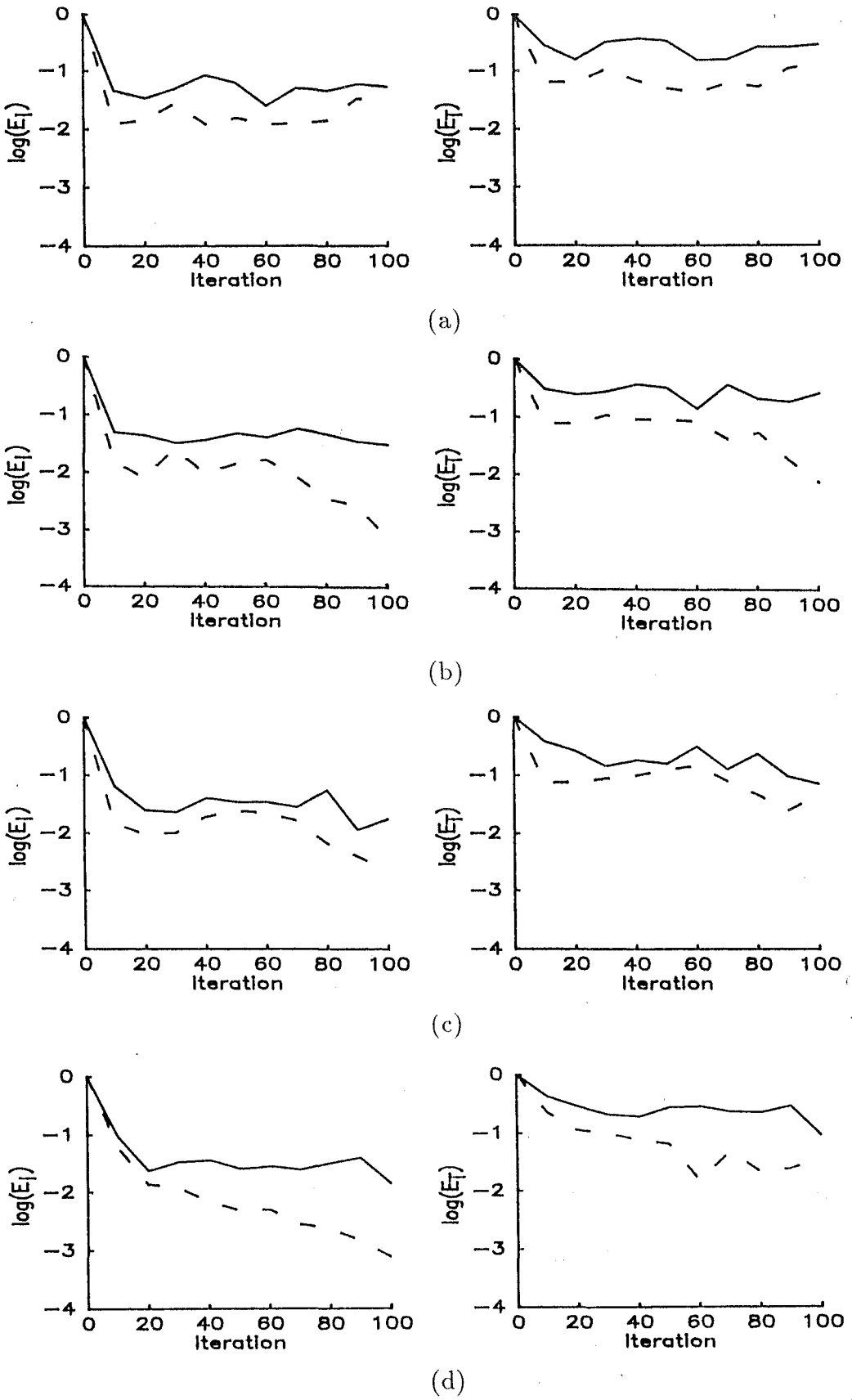


Figure 5.21: Error curves (E_I and E_T) for the reconstructions of image shown in Fig 5.18a. The solid line shows the worst, and the dashed line the best, reconstructions. (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support

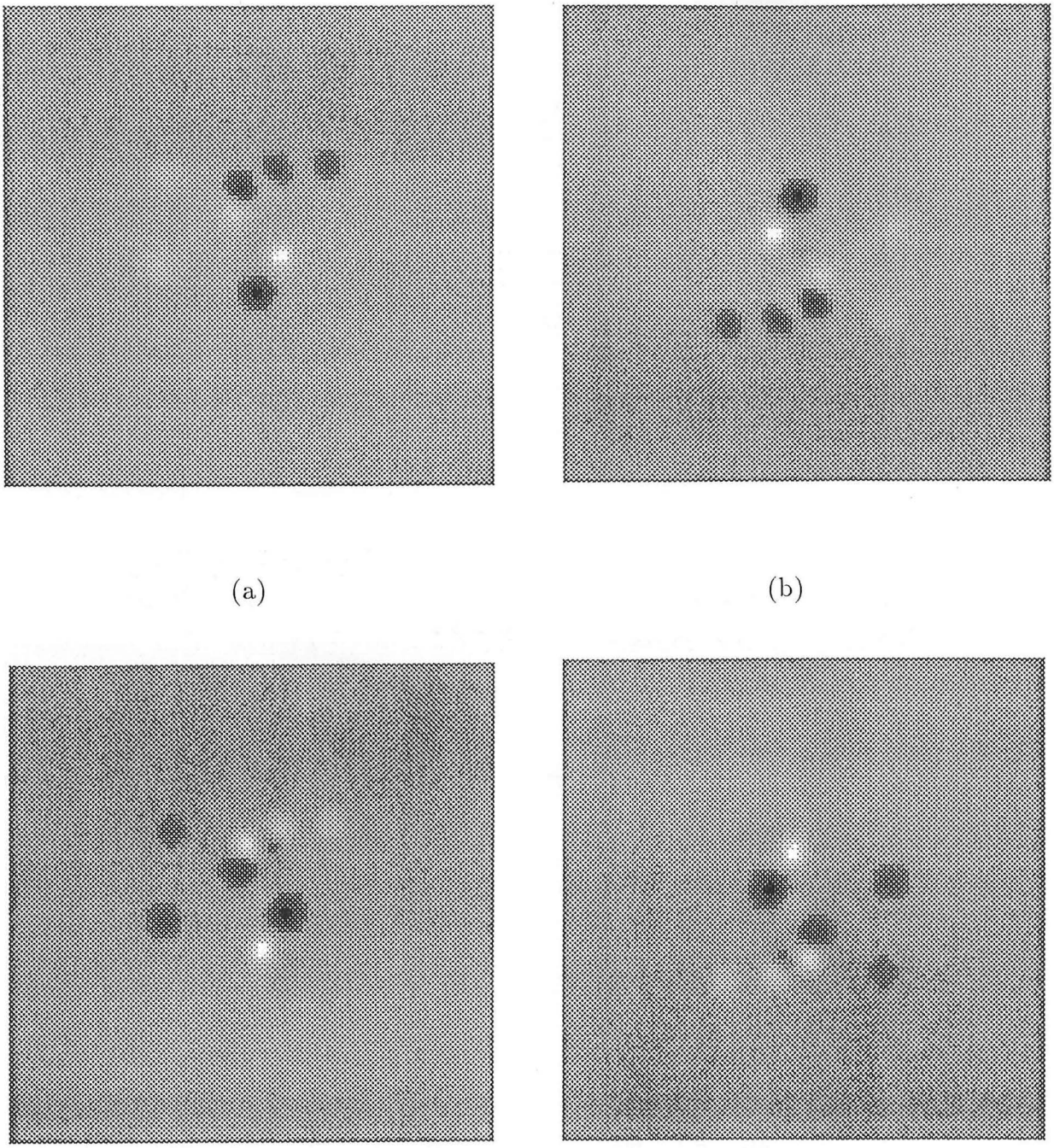


Figure 5.22: Different possible solutions for bipolar image reconstruction. (a) $f(x,y)$ (b) $f(-x,-y)$ (c) $-f(x,y)$ (d) $-f(-x,-y)$.

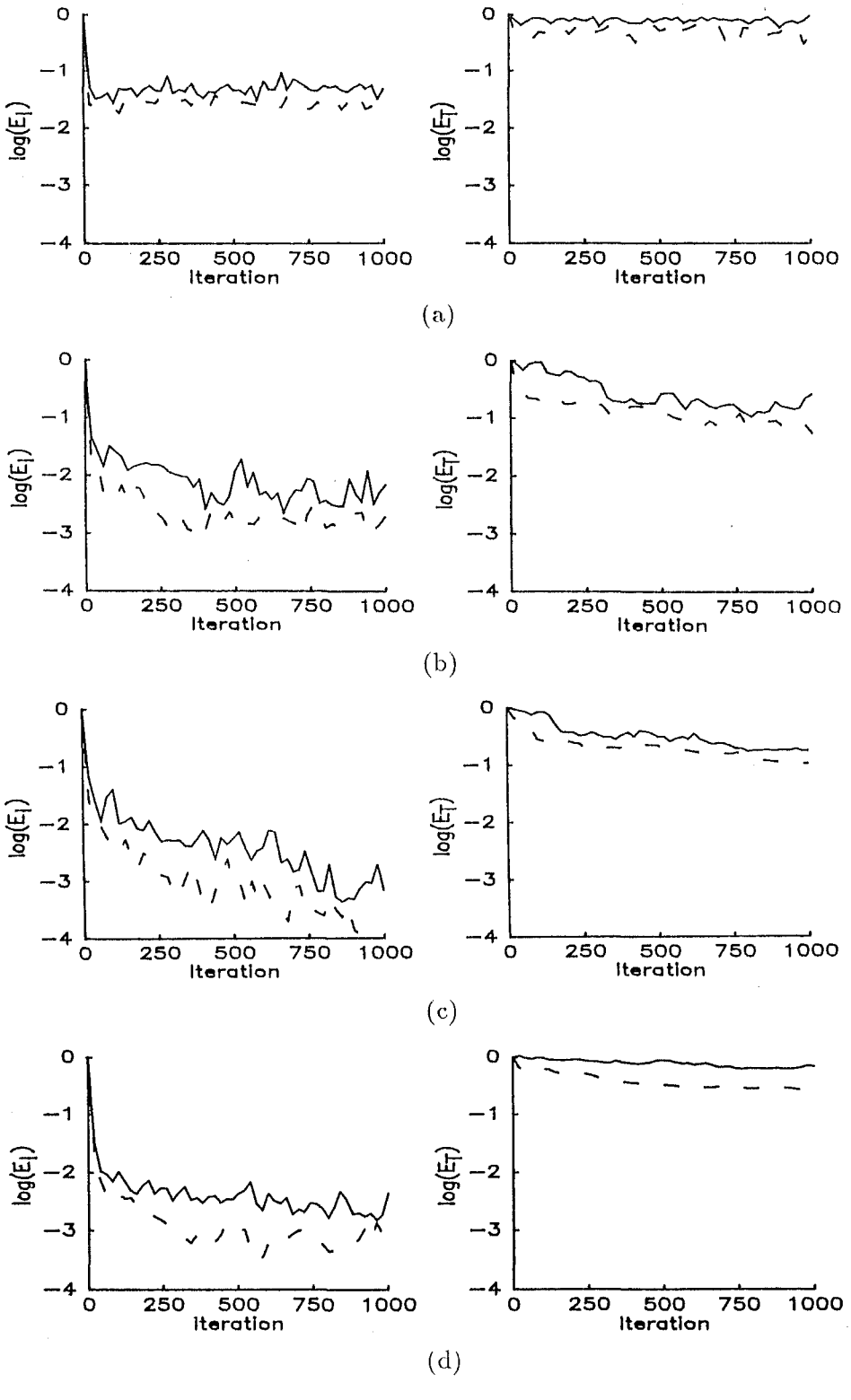


Figure 5.23: Error curves for reconstructions of bipolar image shown in Fig 5.18b. (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support.

Bates 1985). Since the image is no longer constrained to be positive, the image-form can also resemble $-f(x,y)$ (which still possesses the same image-form). There are thus four possible manifestations of the image-form, as shown in Fig 5.22.

It should be noted that the error curves for a bipolar image (Fig 5.23) exhibit a sharper initial decline in E_I compared with the positive images, because negative pixels do not contribute to the E_I of the bipolar image. After several iterations the acceleration in convergence resulting from the added constraint of positivity causes the E_I of the positive image to fall below the E_I of the bipolar images.

In the bipolar reconstructions for a 24 x 24 pixel there is some resemblance to the true object in the best reconstruction (Fig 5.24a). The reconstructions are of generally poor quality since the estimated support is smaller than the true support.

All reconstructions with 28 x 28 and 32 x 32 supports (Figs 5.24b,c and 5.25b,c) resemble one of the expected image-forms, with the best reconstructions being distinguishable from the true image, on a 256 grey level display, only by some background residual error. The 36 x 36 reconstructions, although having some similar features to the true image, are not satisfactory despite seemingly low values of E_I (Fig 5.23d). It appears that, in the absence of a positivity constraint, there is insufficient information available to reconstruct an acceptable image-form when the support constraint is too loose. This is reflected in the high levels of E_T.

It should be noted that significantly more iterations have had to be employed as compared with the positive reconstructions presented earlier. It appears that the reason why Tan and Bates (1985) could not successfully reconstruct bipolar image forms is because they employed insufficient iterations of the less efficient approach of mixing error reduction and hybrid input-output.

Still more iterations are required to recover the complex image. The best and worst reconstructions for the supports tried are shown in Figs 5.26 and 5.27 respectively.

Interestingly, the reconstructions using the 28 x 28 pixel support are better in this case than those using the 32 x 32 pixel support. It appears that, in the absence of constraining the image to be positive or real, a tighter support constraint is an advantage, even though it necessarily results in more truncation of the image-form. The error curves for the complex reconstructions are shown in Fig 5.28.

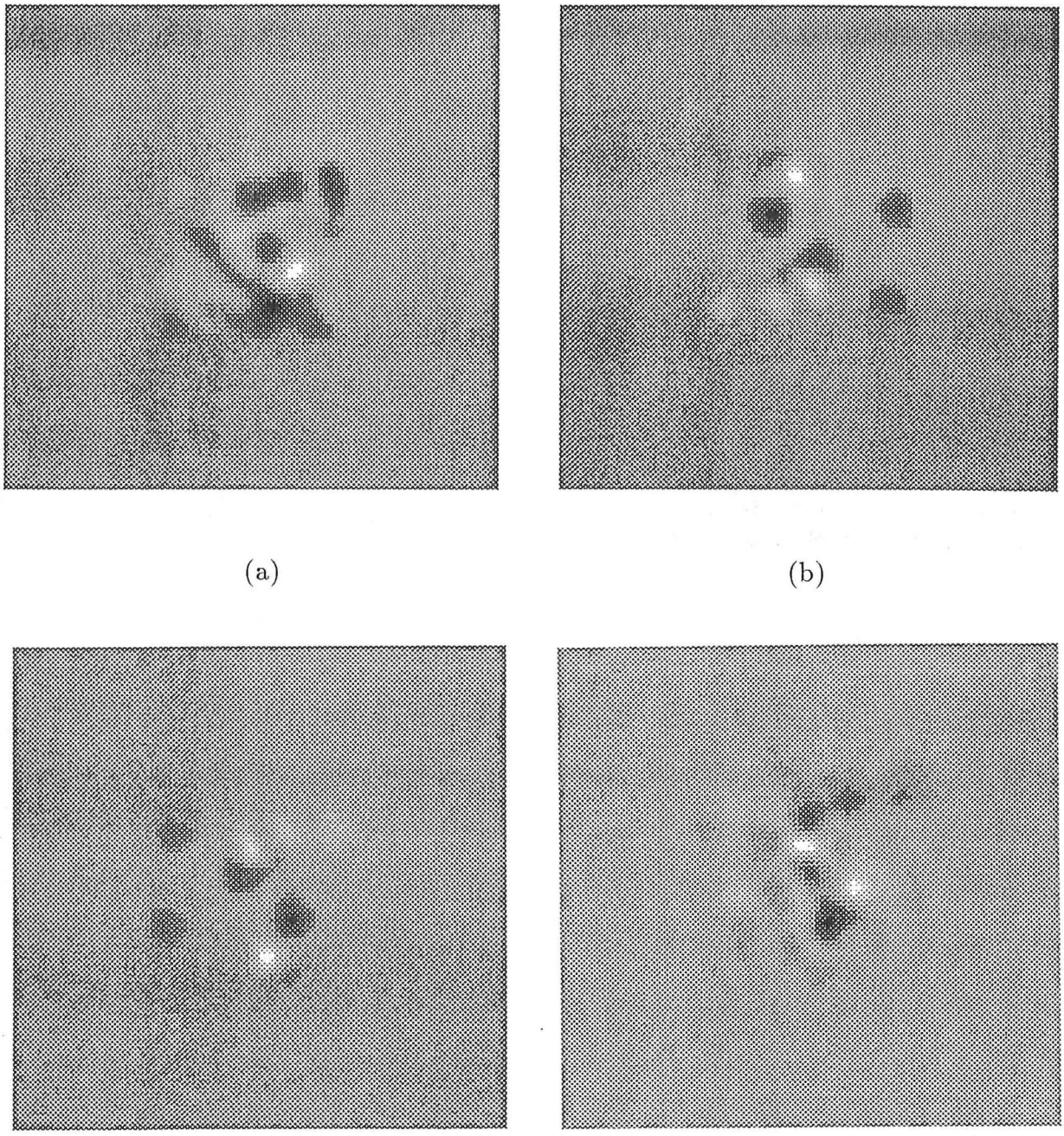


Figure 5.24: Best reconstructions (5 different starting images) of the bipolar image shown in Fig 5.18b, quantised from most negative (black) to most positive (white). (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support.

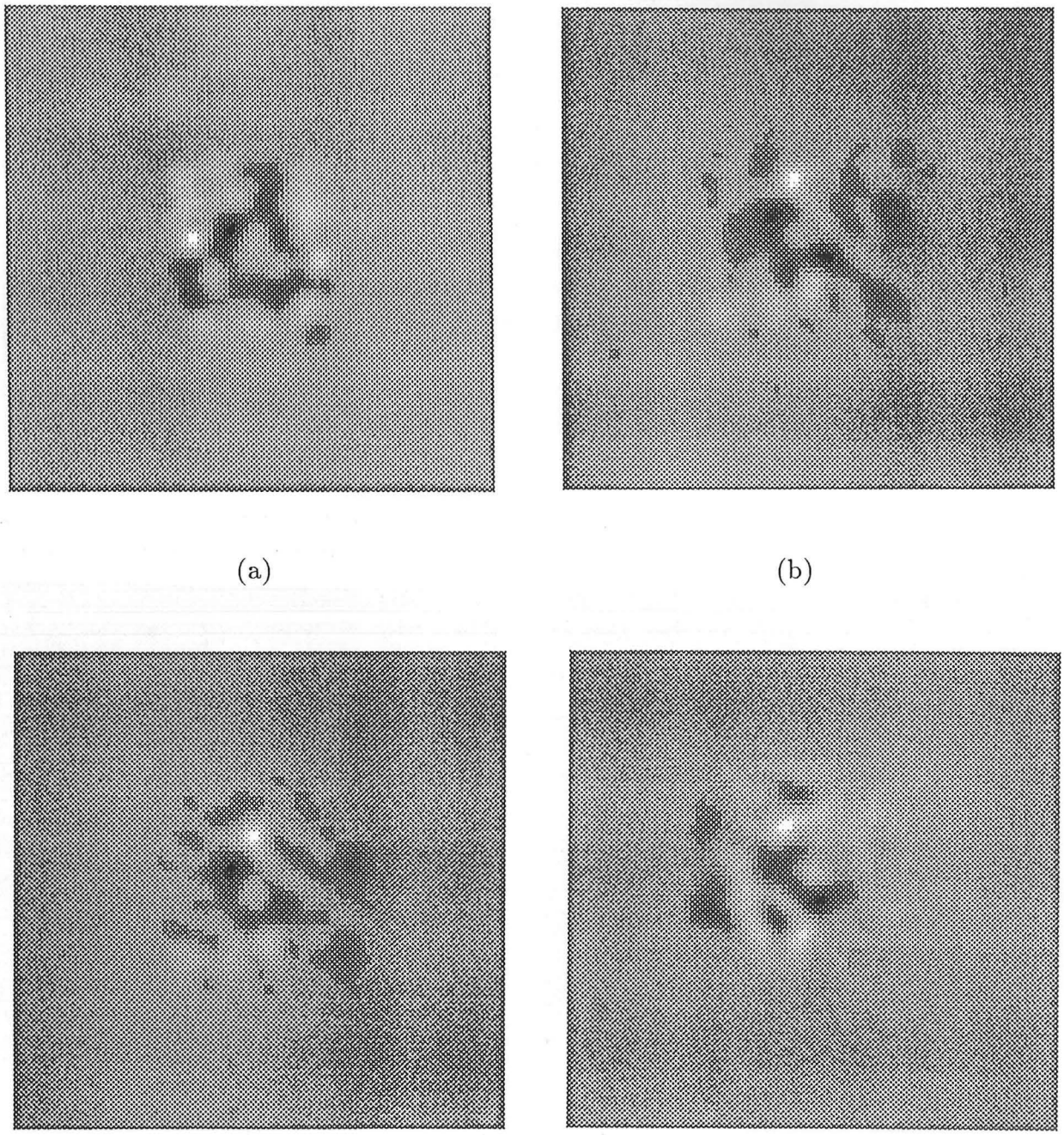
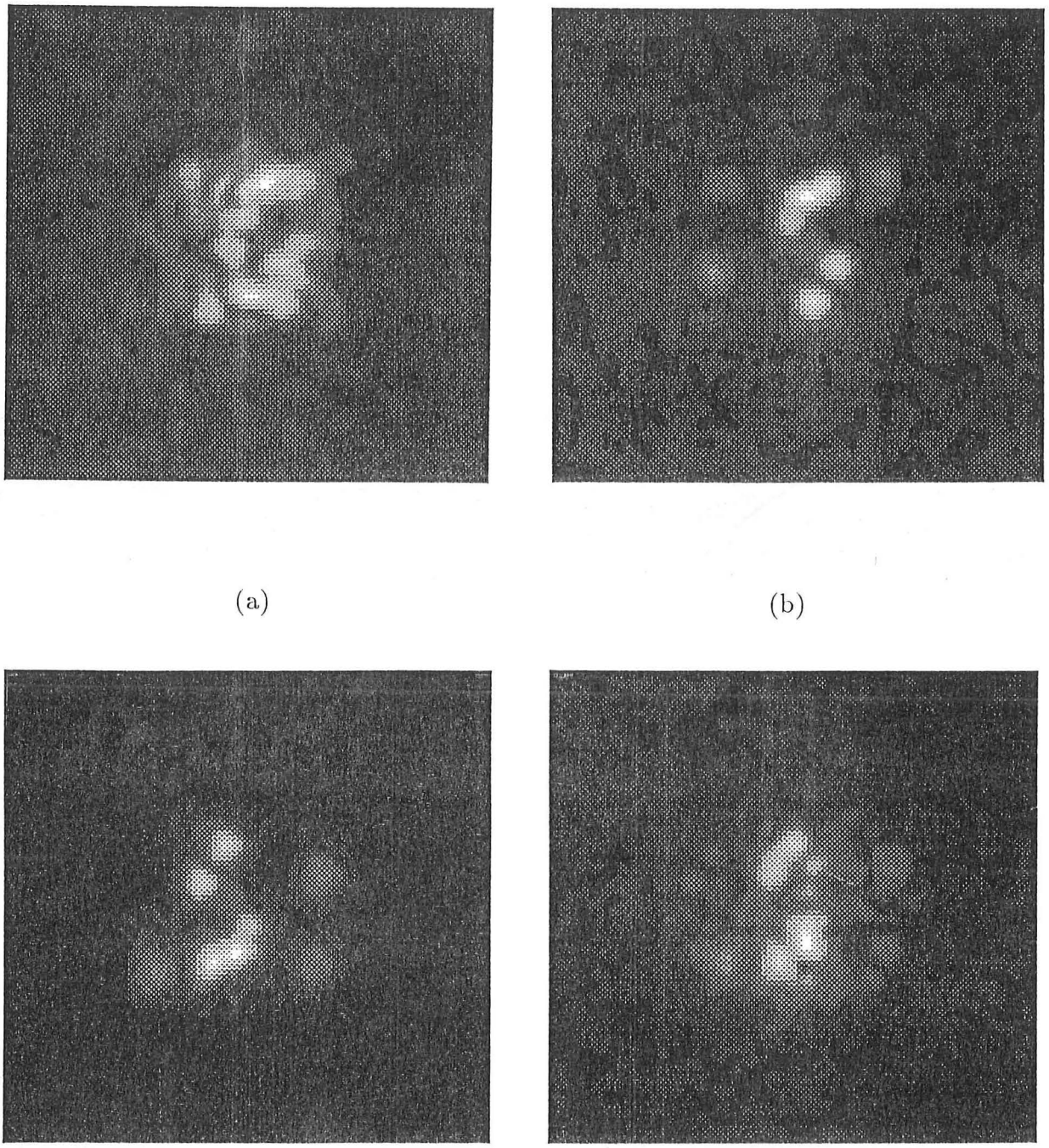


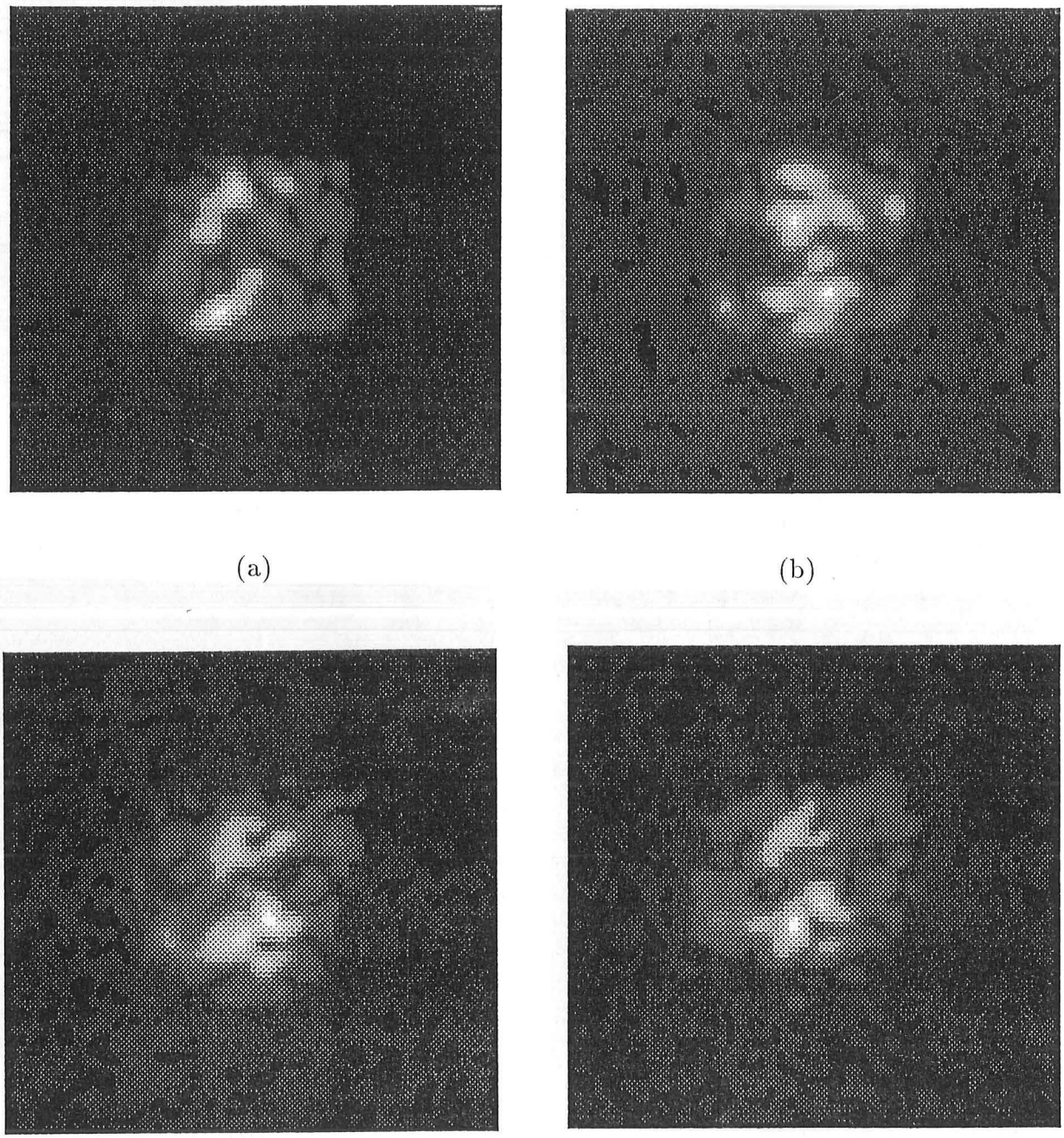
Figure 5.25: Worst reconstructions (5 different starting images) of the bipolar image shown in Fig 5.18b, quantised as in Fig 5.24. (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support



(a) (b)

(c) (d)

Figure 5.26: Best reconstructions (5 different starting guesses) of the complex image shown in Fig 5.18c and d (magnitude only quantised as in Fig 5.8). (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support



(a) (b)

(c) (d)

Figure 5.27: Worst reconstructions (5 different starting guesses) of the complex image shown in Fig 5.18c and d (magnitude only quantised as in Fig 5.8). (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support

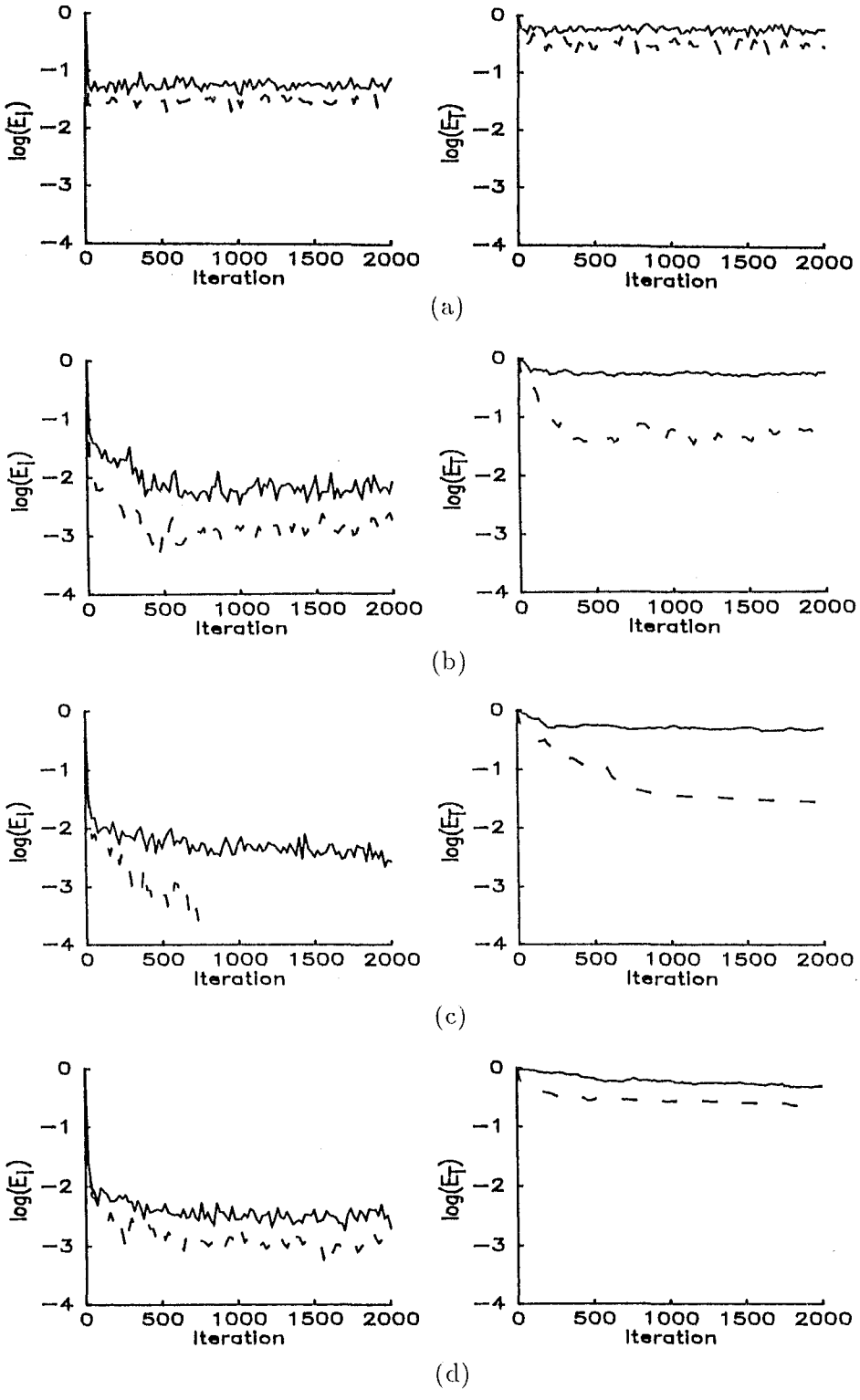


Figure 5.28: Error curves for reconstructions of image shown in Fig 5.18c,d. (a) 24 x 24 pixel support (b) 28 x 28 pixel support (c) 32 x 32 pixel support (d) 36 x 36 pixel support.

5.6 Effects of noise on the Fourier magnitude

The effect of noise on the Fourier phase problem is currently under investigation by McCallum at the University of Canterbury. This section illustrates the robustness of the Fienup iterative algorithms in the presence of noise. The true image used as an example is shown in Fig 5.8a. Noise was added to the Fourier magnitude by generating a pseudo-random bipolar sequence in the range [-m, m], where m is a fraction of $F(0,0)$, which is the average pixel intensity of the image in image space. Since the image was known to be real, symmetry was enforced on the Fourier magnitude by forming the average $|\tilde{F}(u,v)|$, by

$$|\tilde{F}(u,v)| = \frac{1}{2}(|\hat{F}(u,v)| + |\hat{F}(-u,-v)|) \tag{5.18}$$

where $\hat{F}(u,v)$ is the noisy Fourier modulus. $|\tilde{F}(u,v)|$ was then used as an estimate of the true magnitude for Fourier phase retrieval. Since the noise is bipolar there exists the possibility that some of the samples in Fourier space may be negative. All such pixels were set to zero, to ensure that the Fourier magnitude used for reconstructions was positive.

Fraction of F(0,0)	E_N
0.1	2.33 x 10^{0}
0.03	2.36 x 10^{-1}
0.01	3.26 x 10^{-2}
0.003	3.68 x 10^{-3}
0.001	4.40 x 10^{-4}

Table 5.1: Relationship between the noise level as a fraction of F(0,0) and E_N

Table 5.1 lists the values of the ratio E_N, which is defined by

$$E_N = \frac{\int_{(u,v)} |F(u,v)| - |\tilde{F}(u,v)|^2 \, dudv}{\int_{(u,v)} |F(u,v)|^2 \, dudv} \tag{5.19}$$

for the various levels of m. Fig 5.29 illustrates the reconstruction corresponding to the lowest level of E_I which was obtained. It is important to realise that, because the available $|\tilde{F}(u,v)|$ is corrupted, it is not possible to obtain the true image exactly. Because the image has significant fine detail, a visually acceptable reconstruction requires a low level of E_I.

For comparison, Fig 5.30 shows the images obtained when the true Fourier phase is combined with the contaminated Fourier magnitude. Due to the dominance of the phase in determining the image structure (§§2.4 and 2.5) the ideal estimates shown in Fig 5.30 can be expected to provide an upper limit on the quality of the obtainable reconstruction. Although none of the Fienup reconstructions can be said to be faithful replicas of the true image, the appreciable noise levels should be kept in mind. Fig 5.31a shows the image space error for the different levels of noise.

It is important to realise that, when the true phase is added to a noisy Fourier modulus, the resultant estimate of the true image is not compact. It therefore makes

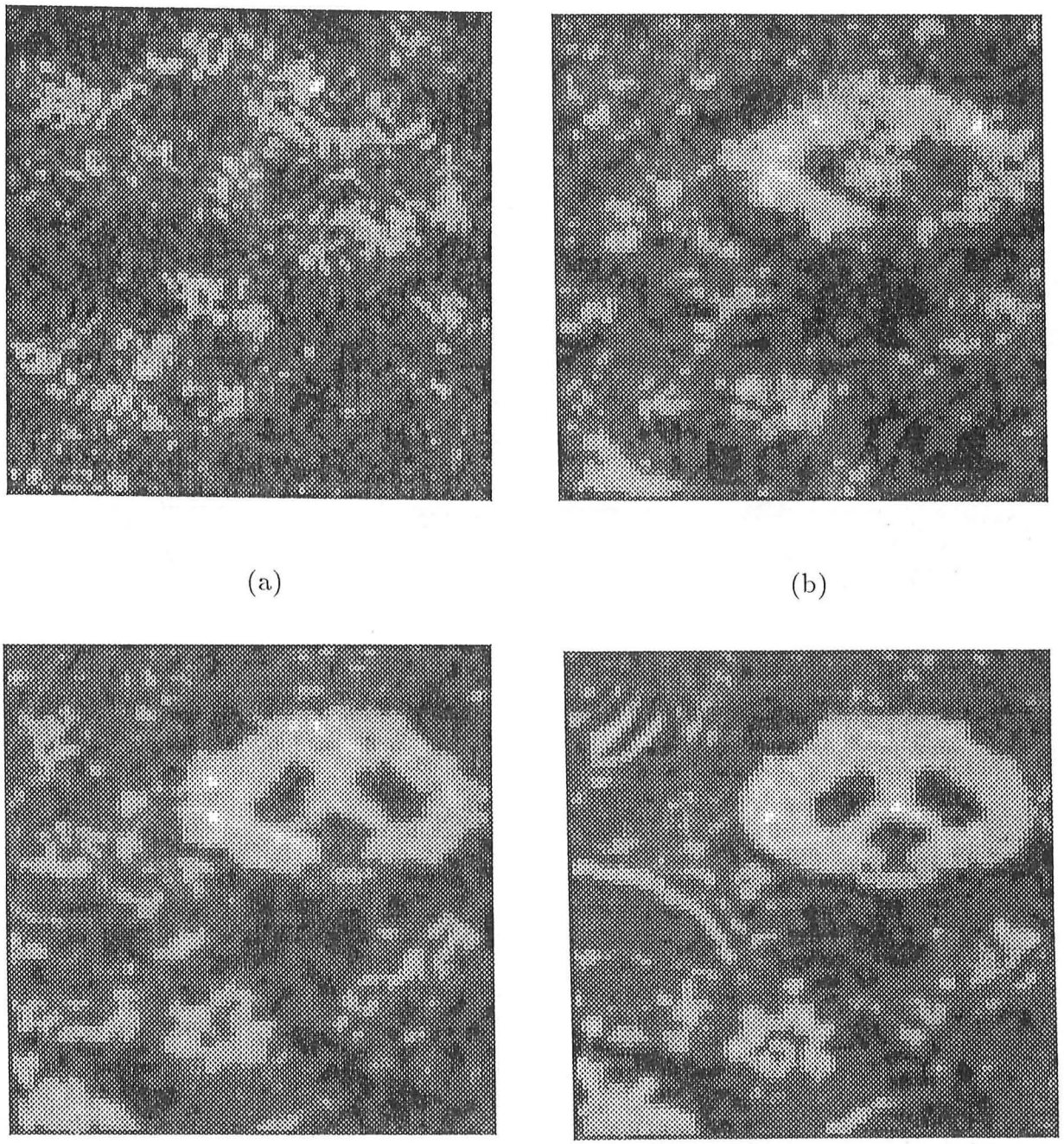


Figure 5.29: Best reconstruction within 500 iterations of image shown in Fig 5.8a. Pseudo-random noise in the range $[-m, m]$ is added where m is a fraction of $|F(0,0)|$. (a) m= 0.03 (b) m= 0.01 (c) m= 0.003 (d) m= 0.001

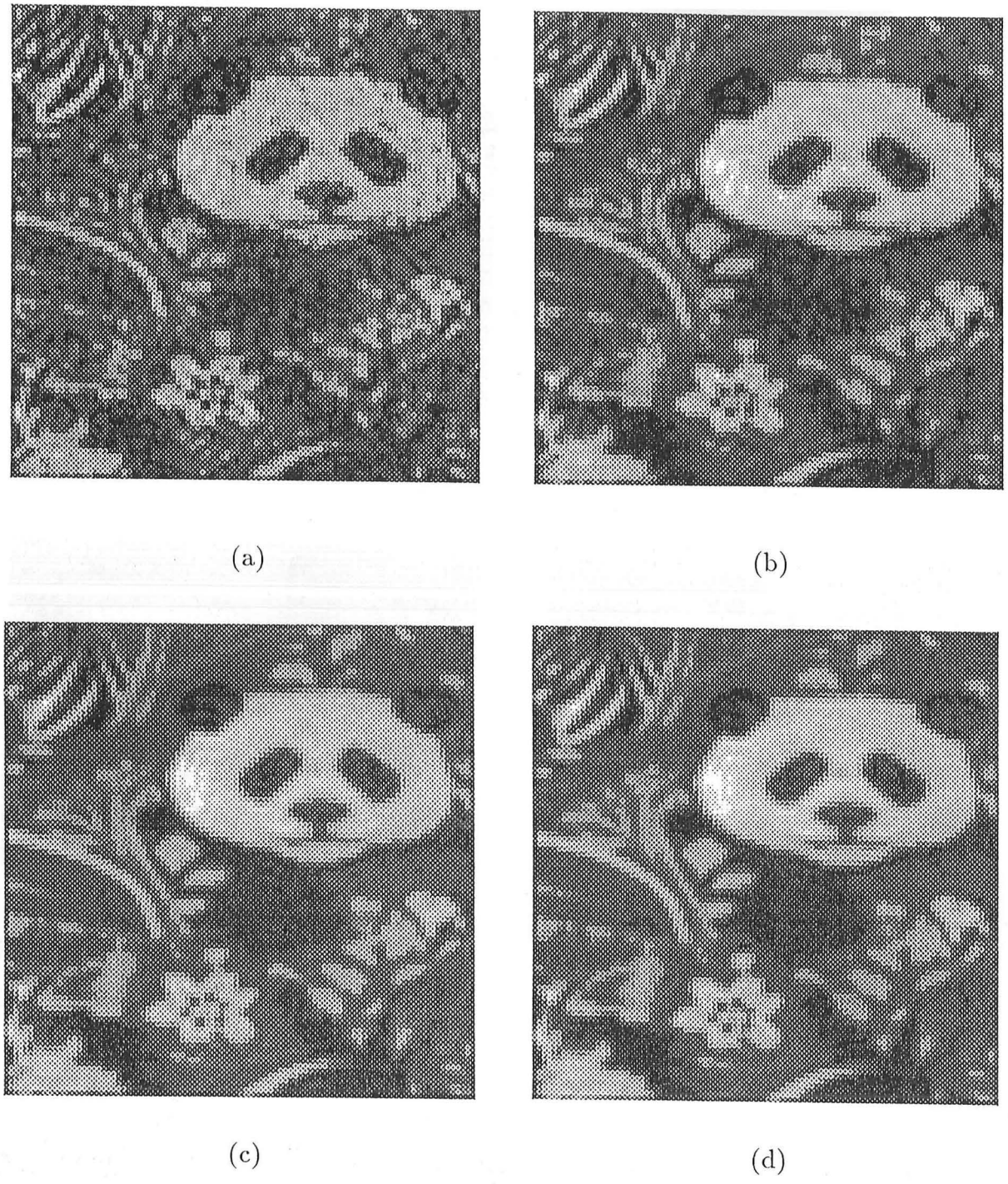
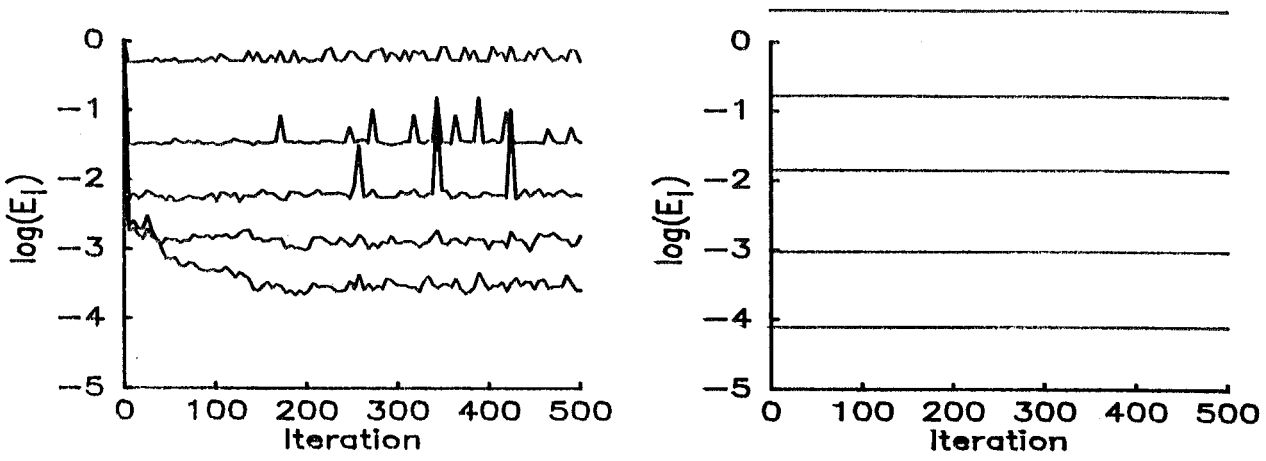


Figure 5.30: Ideal estimates obtained by combining corrupted $|F(u,v)|$ (see Fig 5.29 caption) with correct phase. (a) m= 0.03 (b) m= 0.01 (c) m= 0.003 (d) m= 0.001



(a) (b)

Figure 5.31: Image space error for different levels of noise. Curves are from top to bottom for m=0.1, 0.03, 0.01, 0.003, 0.001. (a) E_I during iteration ($\beta = 0.5$). (b) E_I obtained by combining corrupted Fourier magnitude with correct phase.

sense to calculate E_I for the images shown in Fig 5.30. Fig 5.31b displays such values of E_I. Comparison of Figs 5.31a and 5.31b shows that only for the lowest levels of noise, corresponding to $m = 0.003$ and $m = 0.001$, is E_I lower for the true phase images. Hence the Fienup reconstructions for higher levels of noise are, in fact, better approximations to the available constraints (a point noted previously in Feldkamp and Fienup 1980).

Something which is particularly significant is the relationship of the quality of reconstruction to the level of added noise. Fig 5.32 shows E_T plotted against E_N for the images shown in Figs 5.29 and 5.30. It should be noted that E_N differs from E_F, as defined in (5.6), because it gives a measure of how much the available Fourier modulus is corrupted, because the latter determines how well the estimate agrees with the available Fourier modulus data.

Whilst the Fienup reconstructions exhibit a relationship of the form

$$E_T \propto E_N^{\frac{1}{2}} \tag{5.20}$$

the reconstructions using the correct phase exhibit a relationship of the form

$$E_T \propto E_N \tag{5.21}$$

Thus, to improve the quality of reconstruction by a factor of two, it is necessary to reduce the noise level by a factor of four. This is discussed further in chapter 7.

The behaviour of the reconstruction in the presence of noise is also of interest. Whilst the error curves in Fig 5.29 may appear to have stagnated after a few iterations, this is not in fact so. Although E_I remains at approximately the same level there is a significant difference between the estimate of the image-form before and after an iteration. This feature is currently being used by McCallum at the University of Canterbury to effect significant improvements in the quality of noisy reconstructions (McCallum 1987).

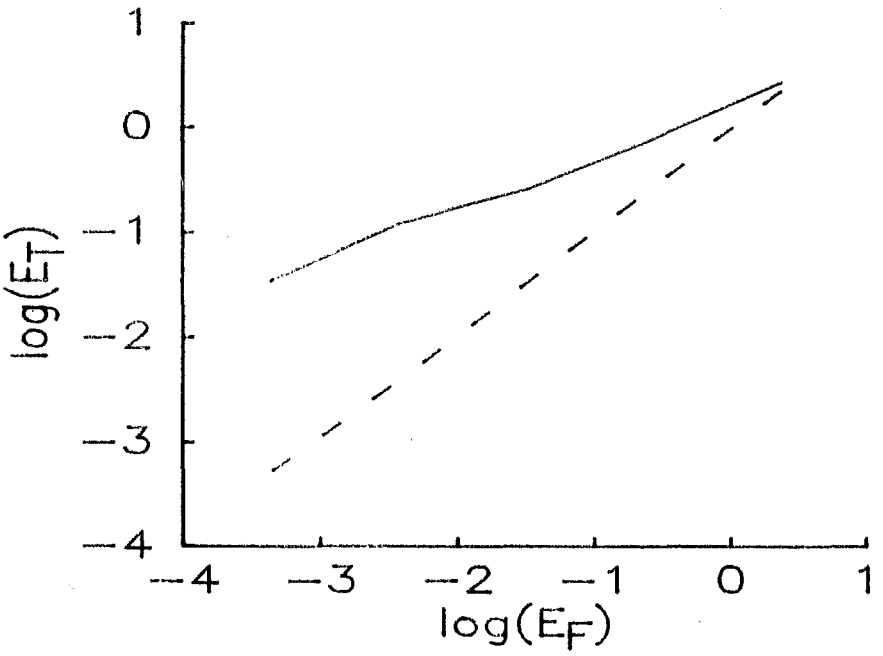


Figure 5.32: E_T for different levels of noise. The solid line shows the E_T obtained from the Fienup reconstructions (Fig 5.29) whilst the dashed line shows the E_T obtained for the reconstructions using the correct phase (Fig 5.30).

5.7 Causes of stagnation in the phase problem

There are a number of reasons for stagnation (Fienup and Wackerman 1986). The basic problem is that, although there is only one image-form, there are a number of possible manifestations of this image-form. The most common form of stagnation encountered when recovering a positive image is convergence to an estimate which resembles a combination of the true image and its reflection in the coordinate origin, as illustrated in Fig 5.33. The image is thus approximately symmetric, and the iteration is therefore stagnant for the reasons given in §5.1. In practice, the estimate is not exactly symmetric and one image eventually becomes dominant, although this may require a large number of iterations. One way of overcoming this type of stagnation is to enforce a smaller asymmetric support for a few iterations (Fienup and Wackerman 1986). Although this temporarily causes divergence from the true image, it does bias the solution towards one of the manifestatins of the image-form. Thus, when the correct support is reimposed, the algorithm usually converges quickly to the correct image-form.

Another form of corruption is stripes appearing on the reconstructed image. The stripes are usually more apparent outside the support as illustrated in the image reconstructed in Fig 5.34. The cause of stripes seems to be the creation of an implicit zero in the Fourier magnitude of the reconstruction which does not correspond to a zero in the Fourier magnitude of the true image (Fienup and Wackerman 1986). This false zero is characterised by a 2π phase shift around a circuit enclosing the false zero in Fourier space (in the manner described in §3.5). Clearly, this zero cannot be explicit, in the sense of occuring at one of the sampling points in Fourier space because at these points the Fourier magnitude is set to that of the true image. The false zero must occur somewhere

Figure 5.33: Stagnation due to convergence to estimate between true image and its reflection in the coordinate origin (compare with Fig 5.8a).

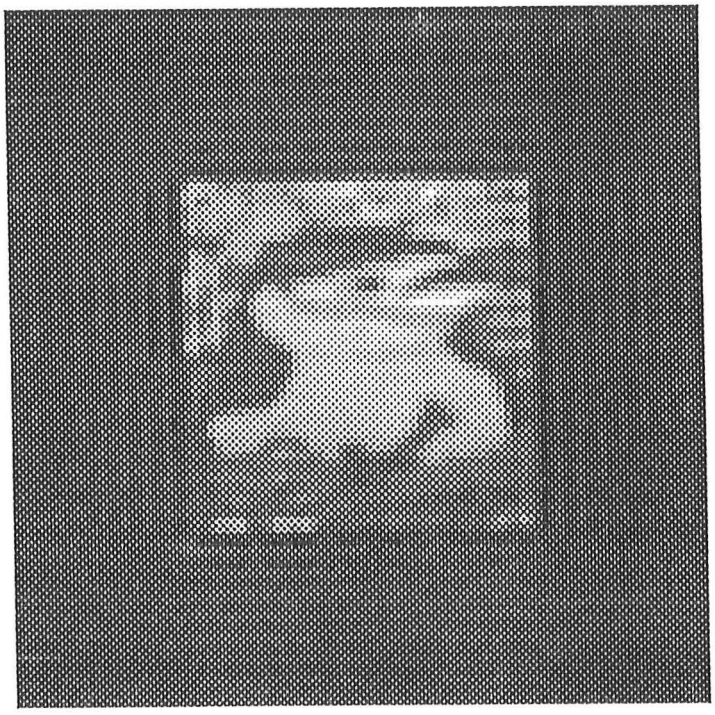


Figure 5.34: Stagnation to a striped image (compare with Fig 5.16).

in between the sample points, implying that it may only cause a small error in the Fourier magnitude samples, even though it results in an appreciable phase error.

The presence of stripes on an image does not usually seriously distort the reconstruction and is in most cases merely an annoyance rather than a serious degradation. A reconstruction with stripes does, however, seem to correspond to a local minimum in E_I. Furthermore, my experience suggests that stripes are less likely to occur when the iterations consist solely of hybrid input-output, rather than a mixture of error-reduction and hybrid input-output.

A final form of stagnation results when the phase recovery algorithm attempts to reconstruct the image-form shifted relative to the support. Even if the bulk of the image-form may lie within the support, there is necessarily some truncation of the image-form at the edges of the support. For objects with tapered edges, there is very little energy contained in this truncated portion of the image-form compared with the portion of the image-form already reconstructed within the support. Hence, the final convergence can be very slow because the change made from one iteration to the next is often mainly determined by E_I.

Some success in overcoming the problem of image-form truncation can be achieved by shifting the support so as to maximise the energy within the support (Fienup and Wackerman 1986, Fright 1984). Alternatively one can try different locations for the support and see which gives the largest reduction in E_I after a small fixed number of iterations. This is discussed further in chapter 7.

Chapter 6

BLIND DECONVOLUTION

This chapter describes techniques of blind deconvolution other than the partitioning of zero-sheets discussed in chapter 4. The first method, which is iterative is described in §§6.1 - 6.3. Unlike conventional blind deconvolutional techniques, which operate on ensembles of blurred images, this method requires only a single blurred image. The second method, which is conventional in that it requires an ensemble of differently blurred images, is a potentially powerful extension of established techniques for the speckle imaging of astronomical objects.

Chapters 4 and 5 note that, for the Fourier phase problem, it is not possible to recover a unique image-form from the Fourier magnitude when the true image is a convolution. Although this non-uniqueness can be a stumbling block in phase retrieval, it can be used to advantage for blind deconvolution. The first deconvolution technique discussed in this chapter hinges upon recovering the multiple image-forms that are associated with the visibility magnitude of a convolution. These image-forms are then combined to derive information about the phases of the components of the convolution. However, for reasons discussed in §6.2, it is only possible to find the phase modulo π. Consequently, in order to determine the true phase at any point (u', v') in Fourier space, it is necessary to determine whether

$$\mathcal{P}[F(u,v)] = \mathcal{P}[F(u,v)] \; mod \; \pi \tag{6.1}$$

or

$$\mathcal{P}[F(u,v)] = (\mathcal{P}[F(u,v)] \; mod \; \pi) + \pi \tag{6.2}$$

The magnitude problem, introduced in §2.5, relates to situations where the magnitude of the true visibility is unknown. The term "pure magnitude problem" is used to describe situations where the true phase is known, whilst the term "modified magnitude problem" is used when the phase is only known modulo π. Although there are many similarities between the pure and modified magnitude problems there are important differences which are discussed below.

Before analysing the modified magnitude problem, it is important to discuss some of the difficulties which occur in iterative solution of the magnitude problem. Although the pure magnitude problem can be formulated as a system of linear equations (Bruck and Sodin 1983; Hayes 1982), this is not possible for the modified magnitude problem. Since the pure magnitude problem is discussed only as an introduction to the recovery from visibility phase modulo π, the linear equations approach is not discussed further herein, but the interested reader is referred to the works cited above and in §6.1.

While there are fewer convergence difficulties for the pure magnitude problem, as

opposed to the pure phase problem, such difficulties can occasionally occur. Unlike for the phase problem, there seems no difficulty with local minima, and slow convergence can be readily overcome by employing quite simple adaptions of the basic iterative loop. §6.1 shows how a simple technique first described by Hayes (1982) provides much improved convergence.

In the Fourier phase problem it is possible to calculate the autocorrelation of the true image. Using the techniques of §5.2 it is then possible to estimate the image-form support from the support of its autocorrelation. There is no corresponding technique of support estimation, however, for the magnitude problem. §6.1 discusses how overestimates of the size of the image support lead to an infinite number of possible solutions for the magnitude problem. In the pure magnitude problem an ambiguous solution is the convolution of the true image with a symmetric image added to a dominant delta function at the origin of image space.

§6.2 deals with the modified magnitude problem. The section starts with a discussion of how modulo π phase is derived from a convolution. The modified magnitude problem also suffers from ambiguities when the support size is made too large. Unfortunately, the ambiguities in the modified magnitude problem are of a more serious nature than those for the pure magnitude problem. Hence an acceptable reconstruction of the true image is only possible when a good estimate of the support is available.

§6.3 deals specifically with problems related to blind deconvolution. It is important to realise that in blind deconvolution there are in fact two interconnected magnitude problems. This can be used to advantage in two ways. Firstly, the two options for the true phase given in (6.1) and (6.2) can not be chosen independently. Secondly, the product of their visibilities' magnitudes is also known and can be used to test whether the components of the convolution have been successfully recovered. Thus, rather than attempt to recover the convolution's components independently from their visibilities' modulo π phases, it is better to reconstruct the image and the psf simultaneously.

The final section (§6.4) deals with a two-dimensional extension of the zero and add technique pioneered in one-dimension by Sinton, Davey and Bates (Sinton et al 1986; Davey et al 1986). Although some astronomical problems are one-dimensional in nature (e.g. the determination of the relative brightness of two stars in a binary system), in general images are two-dimensional. Zero-and-add appears more robust in the two-dimensional case.

6.1 The pure magnitude problem

The computational loop employed to recover the Fourier magnitude is described in §5.1. In the following discussion the notation of §5.1 is employed to describe the iterative solution of the magnitude problem. In addition, $B'_f(x,y)$ is used to denote an estimate of the true image-box $B_f(x,y)$ (cf §1.4).

The major difference between the iterative loops employed to solve the magnitude and phase problems is in the application of the Fourier constraints. In spite of the similarities in their respective iterative loops, the magnitude problem differs significantly from the phase problem in a number of ways. A major difference is that the location and orientation of the true image are contained in the phase of the true visibility. Consequently the terminology of image-form, introduced to describe the solution to the Fourier phase problem, is inappropriate for the magnitude problem. It is, however, only possible to

determine the true image to within a real scale factor from the phase of its visibility. Hence the term normalised image is used in this thesis to describe the solution to the magnitude problem.

Another significant difference is that when the visibility phase is given, rather than the magnitude, there are no theorems for estimating $B'_f(x,y)$. This difficulty is compounded by multiple solutions that occur in the magnitude problem when $B'_f(x,y)$ is larger than $B_f(x,y)$. These multiple solutions are in fact the convolution of the true image $f(x,y)$ and a compact computationally induced psf $h_i(x,y)$, i.e.

$$\tilde{g}(x,y) = f(x,y) \odot h_i(x,y) \tag{6.3}$$

where $\tilde{g}(x,y)$ is used to denote an alternative solution to the magnitude problem. Since $h_i(x,y)$ is, as shown below, derived from a symmetric image it is convenient to introduce the quantity $s(x,y)$ to represent an arbitrary symmetric image.

In §1.5 it is noted that the convolution of two compact images is larger than the individual images. Consequently $\tilde{g}(x,y)$ can still lie within $B'_f(x,y)$, and thus meet the imposed image space constraints, provided $B'_f(x,y)$ is larger than $B_f(x,y)$,

$$B'_f(x,y) \supset B_g(x,y) \supset B_f(x,y) \tag{6.4}$$

The convolution $\tilde{g}(x,y)$ can, however, only be an alternative solution to the phase or magnitude problems when it also meets the Fourier space constraints.

Invoking this notation in the context of the phase problem, it is apparent that $|H_i(u,v)|$ must be unity, since $|F(u,v)|$ must equal $|\tilde{G}(u,v)|$. The requirement that $|H_i(u,v)|$ is unity is, in general, only compatible with the requirement that $h_i(x,y)$ be compact in image space if $h_i(x,y)$ is a delta function. The delta function, although satisfying both the Fourier and image space constraints, does not cause $\tilde{g}(x,y)$ to differ from $f(x,y)$. As a result the image-form is still recoverable when the support is overestimated (as demonstrated in §5.5).

By contrast the Fourier constraints of the pure magnitude problem are met when $\mathcal{P}[G(u,v)] = \mathcal{P}[F(u,v)]$, which requires that $\mathcal{P}[H_i(u,v)] = 0$. This constraint on $\mathcal{P}[H_i(u,v)]$ can be satisfied by forming $h_i(x,y)$ from a symmetric image $s(x,y)$ plus a delta function of sufficient amplitude at the origin of image space. Thus,

$$h_i(x,y) = s(x,y) + A\delta(0,0) \tag{6.5}$$

is always zero phase provided A is a suitably large (relative to the maximum magnitude of $s(x,y)$) positive constant. Thus the magnitude problem can justifiably be considered more ambiguous than the phase problem (Bates and Lane 1987, SPIE).

Although a number of authors (Hayes 1982; Bruck and Sodin 1983; Levi and Stark 1984) have noted that $h_i(x,y)$ must be symmetric, there are two important points that need to be emphasised concerning the use of iterative techniques for magnitude recovery. Firstly not all symmetric images generate ambiguous solutions to the pure magnitude problem because the requirement of zero phase in Fourier space is significantly stricter than the requirement that an image is symmetric in image space. It should be noted that, since convolution is equivalent to multiplication in Fourier (or Z) space, convolutions which have a symmetric component are often referred to as having symmetric factors (Hayes 1982).

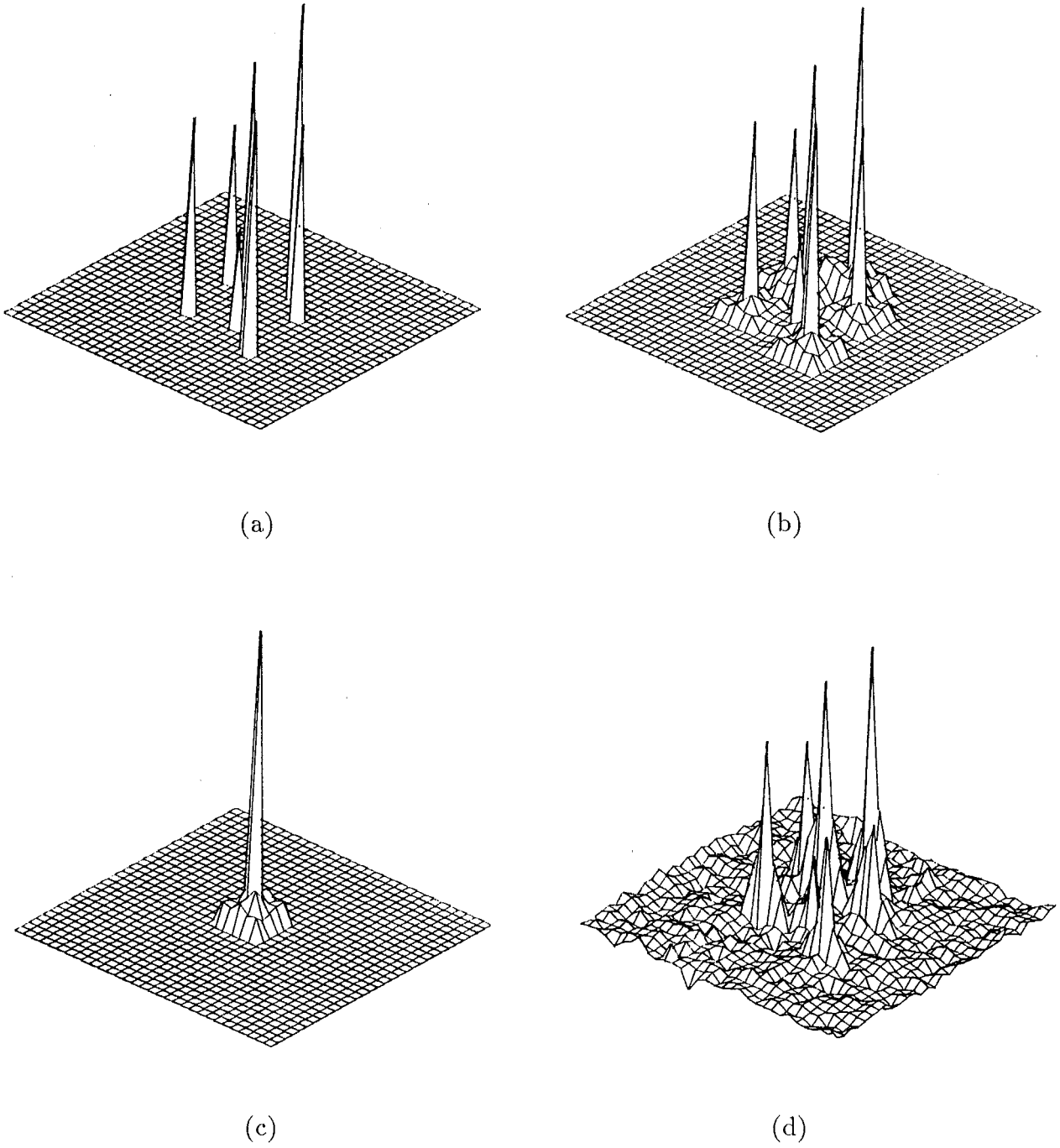
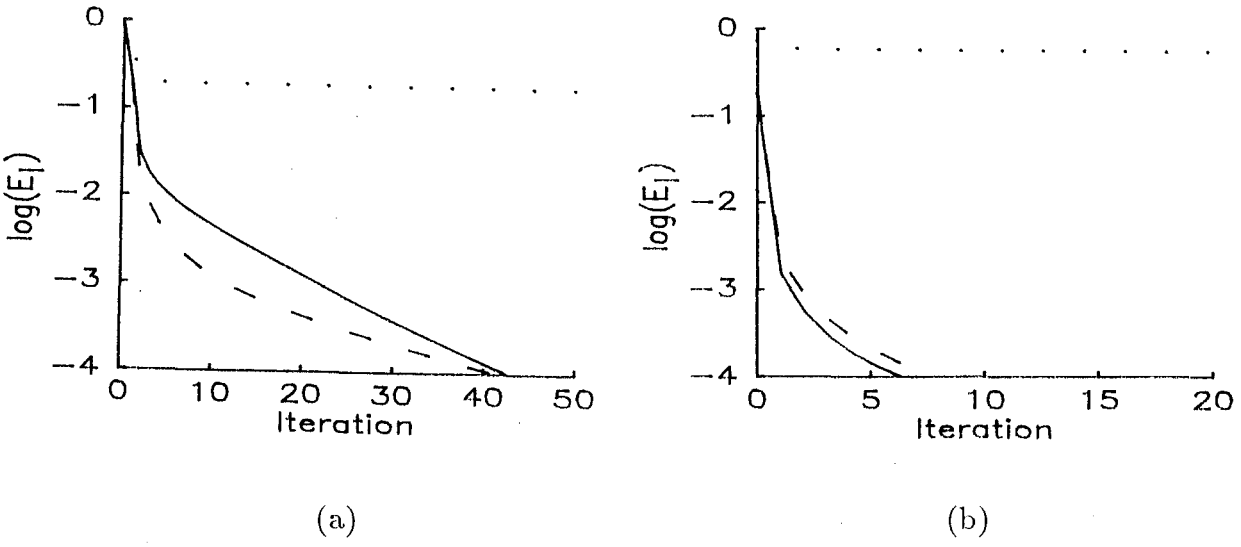


Figure 6.1: Effect of incorrect estimation of $B'_f(x,y)$ on reconstructions from the Fourier phase. (a) The true image $f(x,y)$ (b) The reconstruction with $B'_f(x,y)$ too large. (c) The symmetric blurring function $h(x,y)$ ($\mathcal{P}[H(u,v)] = 0$). (d) The reconstruction with $B'_f(x,y)$ too small.



(a) (b)

Figure 6.2: E_I for the reconstructions from the visibility phase of Fig 6.1a. (- - -) Support too large, (——) Exact support, (.....) Support too small
(a) Random starting magnitude (b) Starting magnitude set equal to a constant

Secondly, although polynomials with symmetric factors form a set of measure zero amongst that the set of polynomials of a given order, this does not mean they do not occur in practice. Using iterative magnitude recovery with too large a support constraint nearly always results in convergence to $\tilde{g}(x,y)$ rather than $f(x,y)$. Only when the estimate of the support is exact is it possible to uniquely determine a solution to the magnitude problem (Bates and Lane 1987, Pfefferkorn).

The occurrence of these computationally induced psfs is most easily seen when dealing with an image comprised of discrete points, such as shown in Fig 6.1a. A reconstruction from the Fourier phase using an iterative loop with a correct estimate of $B_f(x,y)$ is identical to Fig 6.1a. When the support is enlarged convergence again occurs, i.e. the reconstruction again agrees with the constraints in Fourier and image space, but note that the result (Fig 6.1b) is a blurred version of Fig 6.1a. The blurring function is shown in Fig 6.1c, and is a compact image whose visibility phase is identically zero. When too small a support is employed the image shown in Fig 6.1d is reconstructed which fails to satisfy both the Fourier and image space constraints, as witnessed by the high residual E_I shown in the error curves of Fig 6.2.

In my experience, employing an iterative solution to the magnitude problem results in a reconstruction that always fills the extent of the estimated support. In other words, whenever the support is overestimated, the reconstruction is always a blurred version of the true image. The dominance of the delta function ensures that this blurring is not severe. This can be seen from (6.5), since

$$\begin{aligned} g(x,y) &= f(x,y) \odot (s(x,y) + A\delta(0,0)) \\ &= A\, f(x,y) + (f(x,y) \odot s(x,y)) \end{aligned} \tag{6.6}$$

Because A is large, as it must be to ensure zero phase in Fourier space, the reconstruction consists of the true image plus a much lower intensity blurred version of the image. Fig

6.3 illustrates the occurence of symmetric blurring in a more complicated 64 x 64 pixel image. The reconstruction (Fig 6.3b) is again a blurred version of the true image (6.3a) with $h_i(x,y)$ shown in Fig 6.3c. The reconstruction quality is still quite good (E_T in this case = 0.03).

Analysis of the iterative loop for the magnitude problem appears to guarantee convergence (Youla and Webb 1982). The convergence, however, is often slow when dealing with large or complex images. Fortunately, there does not appear to be an analogue of the stagnation that arises in the phase problem where the constraints in image and Fourier space counteract each other. Consequently, in the magnitude problem more sophisticated minimisation techniques are useful in accelerating the minimisation of E_I.

A simple method of improving convergence which is described by Hayes et al (1980), relies on optimising the step size taken in the iterative loop (cf the discussion of (5.13) in the context of the phase problem). Fortunately in the magnitude problem there is an "optimal" (Oppenheim et al 1980) method of calculating Λ without recourse to a line search (a technique described in §5.2), i.e.

$$\Lambda = \frac{\int_{(x,y)} -\Delta k_i(x,y) k_i'(x,y)}{\int_{(x,y)} |\Delta k_i(x,y)|^2} \tag{6.7}$$

Unlike the basic iterative loop (§5.1) which fixes $\Lambda = 1$, choosing Λ according to (6.7) does not guarantee that E_I is monotonically decreasing, unless Λ is restricted to the interval [0,1]. Divergence of the iterative loop has been observed when Λ is unrestricted (Levi and Stark 1984). In practice however convergence is usually much faster when Λ is allowed to assume whatever value is dictated by (6.7).

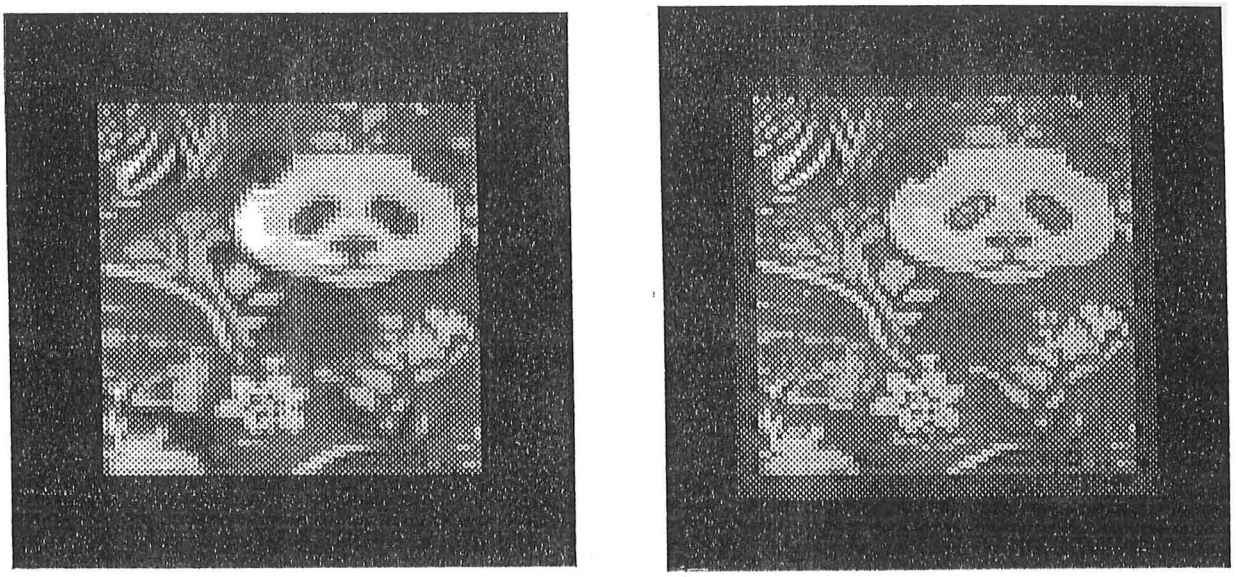
An example of the improved convergence obtained by using an optimal step size is shown in Fig 6.4. The starting image used for both algorithms was formed by combining the true phase with unit magnitude. Clearly a worthwhile improvement in convergence is obtained by employing (6.7).

There are other methods of accelerating convergence in the magnitude problem, for example the technique proposed by Levi and Stark (1984). For the purposes of this thesis, however, the conventional magnitude recovery algorithm, or Hayes' extension of it, has proved adequate. Thus although further modifications to the algorithm could possibly prove worthwhile, they have not been investigated further.

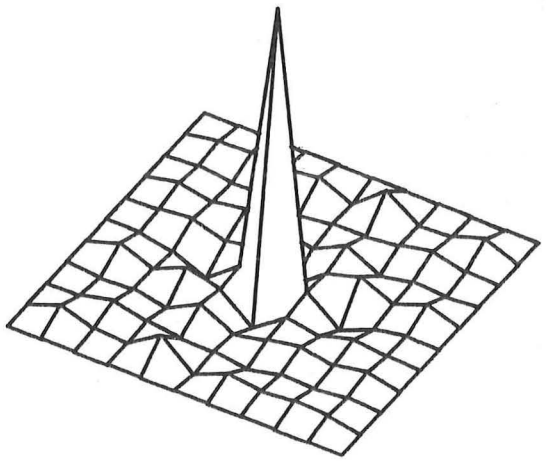
6.2 The modified magnitude problem

Recovery of the normalised image from the Fourier phase modulo π has a number of similarities with the conventional magnitude problem and again the basic iterative loop described in §5.1 is employed. It is apparent that the application of the Fourier space constraint must be modified to select whether (6.1) or (6.2) is chosen as the true phase. The simplest approach is to assume that the true phase is given by whichever of (6.1) or (6.2) is the "closest" to $\mathcal{P}[K_i(u,v)]$ (cf §5.1).

Determining the phase difference between two phases is more difficult than it may appear because in practice all phases are computed, rather than analytically determined,



(a) (b)



(c)

Figure 6.3: Illustration of the effects of too large a support constraint on recovery from the Fourier phase. (a) The true image 64 x 64 pixel image $f(x, y)$ quantised to 32 grey levels ranging from 0 (black) to a normalised maximum of 1 (white) (b) The reconstruction with the 72 x 72 pixel support. (c) The computationally induced psf, $h_i(x, y)$ ($\mathcal{P}[H(u, v)] = 0$).

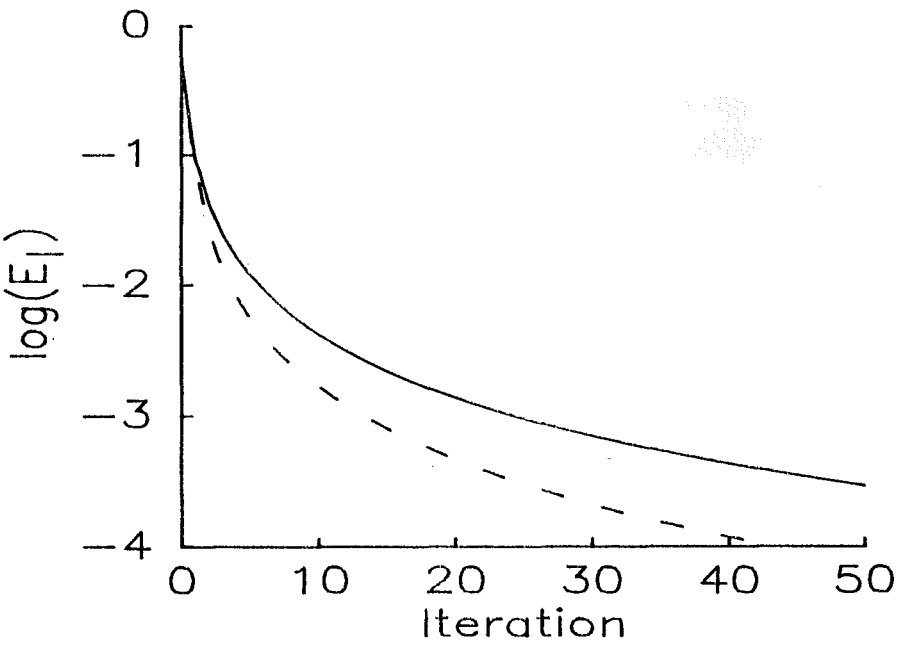


Figure 6.4: Error curves for the reconstruction of the image shown in Fig 6.3a. (——) Using error reduction form of image space constraint. (- - -) Using Hayes' modification of the image space constraint.

and hence the absolute phase is only known modulo 2π. Consider the two phases A and B, where $0 < A < \pi$ and $-\pi < B < 0$ as shown in Fig 6.5. It is apparent that in modulo 2π arithmetic

$$\Theta_1 = |(A - B) \; mod \; 2\pi| \tag{6.8}$$

does not in fact equal

$$\Theta_2 = |(B - A) \; mod \; 2\pi| \tag{6.9}$$

As a result $\mathcal{P}_D[A, B]$, the phase difference between A and B, is defined to be equal to the smaller of Θ_1 and Θ_2. Since Θ_1 and Θ_2 are both positive and $\Theta_1 + \Theta_2 = 2\pi$, $\mathcal{P}_D[A, B]$ can never exceed π.

The most important difference between the pure magnitude and the modified magnitude problems is in the ambiguities which arise when the estimated support is too large. In the pure magnitude problem the symmetric factors in Fourier space are required to be of zero phase. On the other hand, in the modified problem $h_i(x, y)$ need only be symmetric to avoid violating the Fourier space constraints. Since $h_i(x, y)$ need no longer have a dominant point at the origin of image space the blurring is necessarily more severe. This is readily apparent in the recovery of the image shown in Fig 6.3 from its visibility phase modulo π, Fig 6.6a. The symmetric blurring factor is shown in Fig 6.6b. Very much worse distortion is apparent in the reconstruction from the modulo π phase (Fig 6.6a) than for when the true phase is in fact known (Fig 6.3b). This increased distortion is reflected in the much higher level of E_T for the modified magnitude problem ($E_T = 0.3$) when compared with the pure magnitude problem (where $E_T = 0.03$).

One possible method of recovering the magnitude is to find the minimum support for which convergence occurs (Lane and Bates 1987). In the absence of noise E_I declines rapidly with each iteration, for all $B'_f(x, y) \supset B_f(x, y)$, until it stagnates at a level determined by the numerical accuracy of the computer.

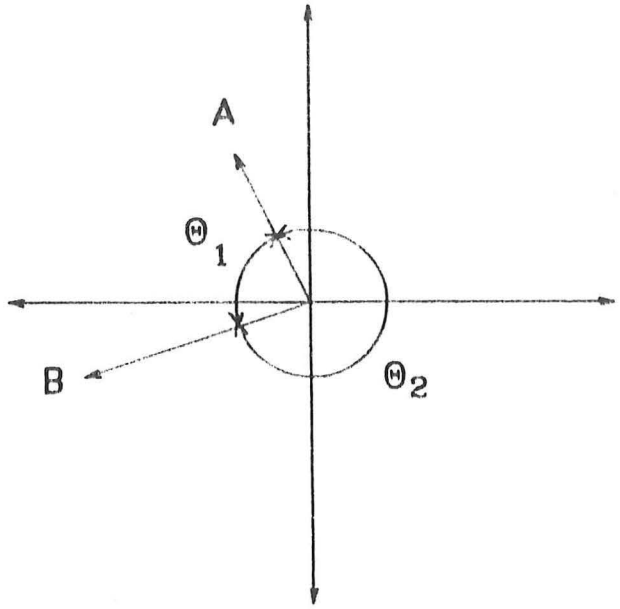
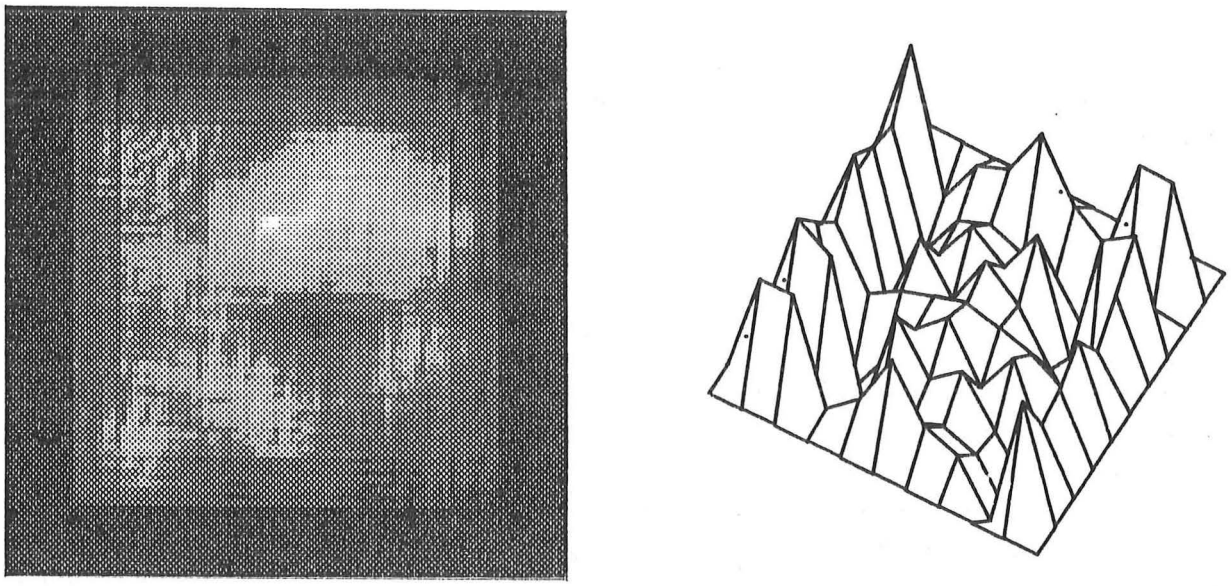


Figure 6.5: Determination of the phase difference between two vectors A and B



(a) (b)

Figure 6.6: Reconstruction of Fig 6.3a from its visibility phase modulo π. (a) Reconstruction within 72 x 72 pixel support, quantised as for Fig 6.3a. (b) Computationally induced psf, $h_i(x, y)$.

Since contamination is always present in the real world, the effects of noise on convergence must be considered, if the convergence of the iterative loop is to be used to differentiate between too large and too small a support. To this end noise was added to the visibility of the image shown in Fig 6.1a. The noise was a symmetric array of pseudo-random numbers (the array was constrained to be symmetric because the image is a priori known to be real). The level of noise is given by N_F, where

$$N_F = \frac{\int_{(u,v)} |C(u,v)|^2}{\int_{(u,v)} |F(u,v)|^2} \tag{6.10}$$

Clearly the noisy Fourier phase is incompatible with an object of compact support and consequently there is a residual E_I when an image is recovered from the modulo π phase. Fig 6.7 illustrates how the distinction between the correct and too small a support becomes less defined with increasing levels of noise. Although for the image used in this example it is still possible, at high levels of N_F, to determine the correct value of $B'_f(x,y)$ from the final values of E_I this method may not be as effective for more complicated images. Fortunately when dealing with blind deconvolution there exists the more exact means of support determination introduced in §6.3. Despites the defects of the approach described in this section the quality of the reconstructions from the noisy modulo π phase (Fig 6.8) indicates that reconstructions are robust in the presence of noise.

Minimising the size of the support also poses difficulties when the true image is a convolution of symmetric and assymmetric components, denoted by $f_a(x,y)$ and $f_s(x,y)$ respectively. Since there is no way for the iterative loop to differentiate between $f_s(x,y)$ and the computationally induced psf $h_i(x,y)$, minimising the support results in the recovery of $f_a(x,y)$. An extreme example is when $f(x,y)$ is symmetric, whereupon minimisation of the support results in the recovery of a delta function. Since

$$\mathcal{P}[S(u,v)] \; mod \; \pi = 0 \tag{6.11}$$

it is not possible to recover a symmetric image from its modulo π phase.

The positioning of the support can also cause difficulties in the magnitude problem. Unlike the image-form in the phase problem the normalised image in the magnitude problem is not invariant under translation. This is because translation of an image by a vector (x_0, y_0) in image space is equivalent to introducing, at all points (u,v) in Fourier space, a phase shift equal to $2\pi(x_0u + y_0v)$. Consequently, in the magnitude problem, unless $B'_f(x,y)$ is correctly positioned convergence to the true image does not occur.

When dealing with sampled images it is possible that this linear phase shift can correspond to shifting an image in image space a distance corresponding to a fraction of a pixel. This introduces an interesting problem in determining the support, especially for images consisting of discrete points such as Fig 6.1a. When the phase in Fourier space is linearly shifted, by an amount equivalent to a translation of half a pixel in image space, it is no longer possible to reconstruct the true image exactly. Minimisation of the support, as described above, results in Fig 6.9a, a blurred version of the true image, since it is not possible to reduce $B'_f(x,y)$ by a fraction of a pixel. The resultant reconstruction is in fact

$$\tilde{g}(x,y) = f\left(x - \frac{\varepsilon_x}{2}, y\right) \odot \left(\delta\left(\frac{\varepsilon_x}{2}, 0\right) + \delta\left(\frac{-\varepsilon_x}{2}, 0\right)\right) \tag{6.12}$$

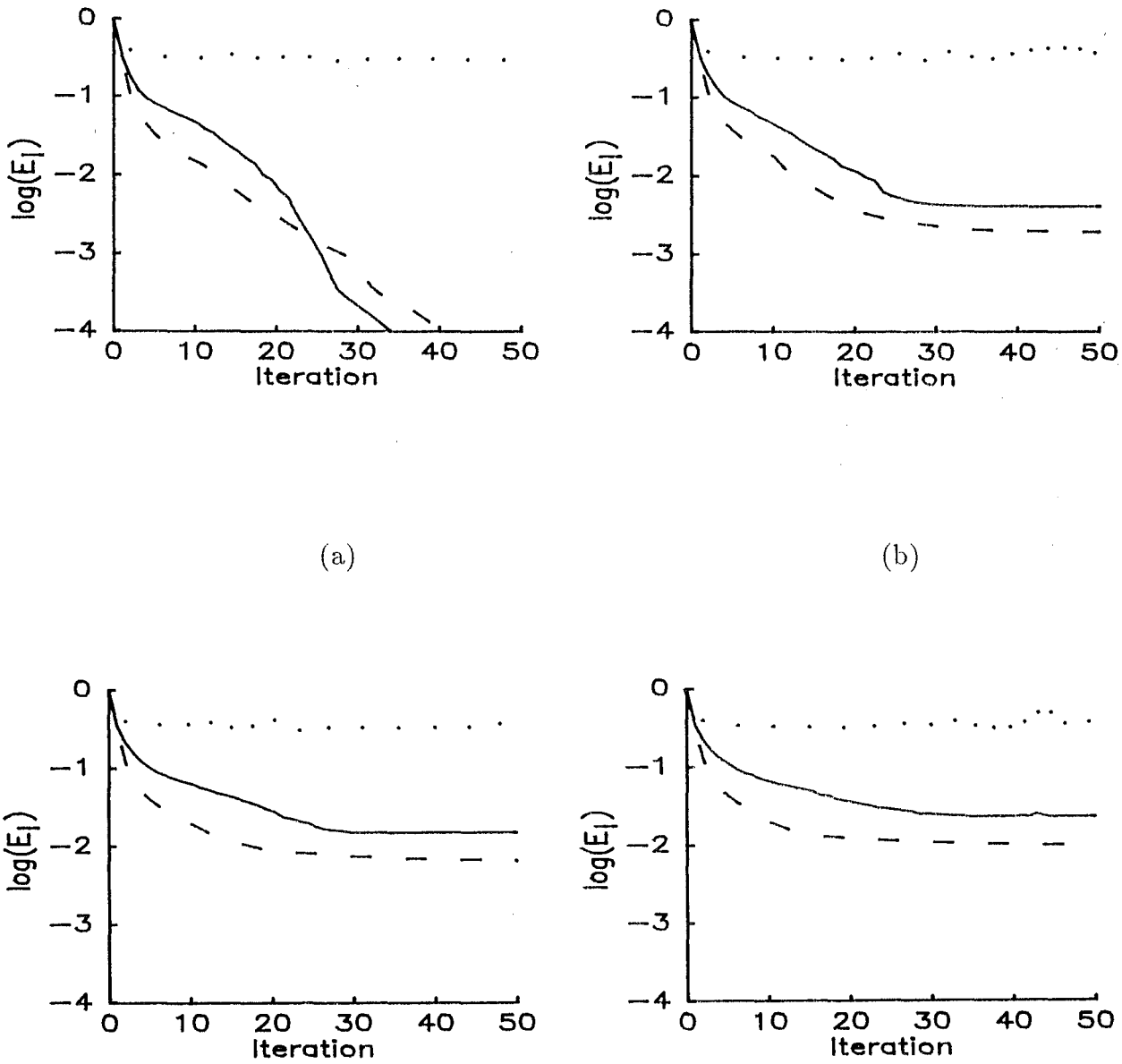
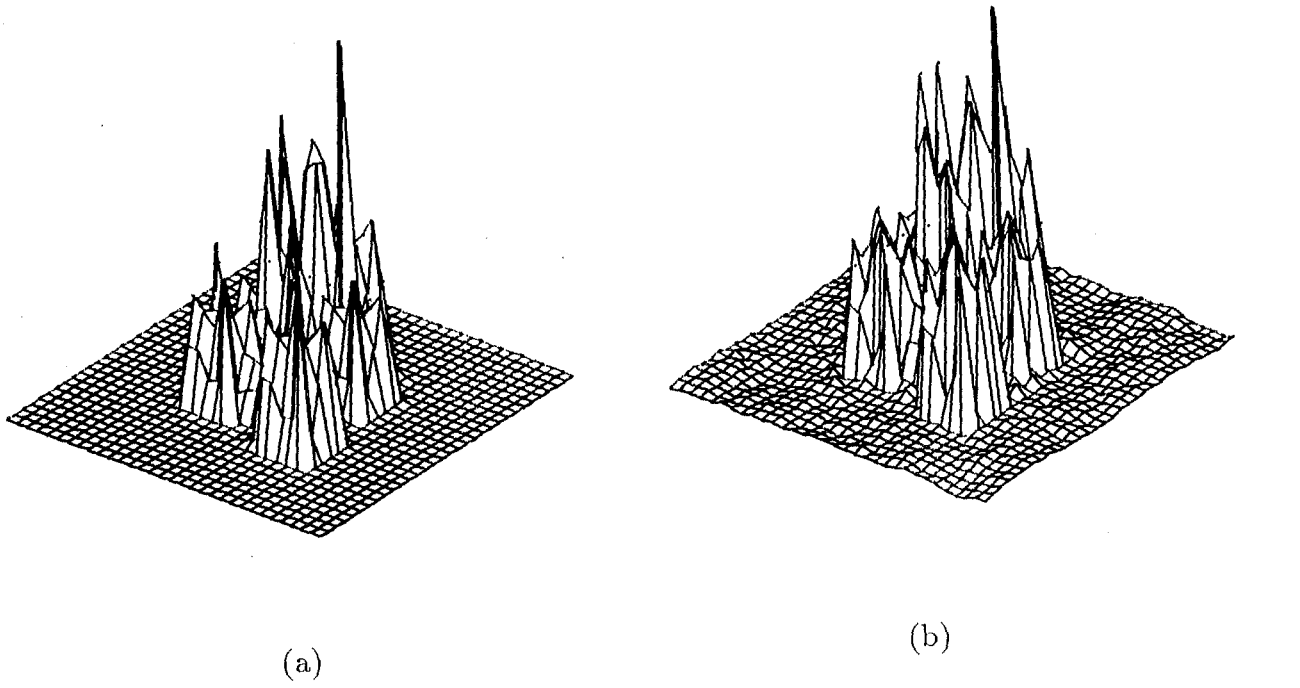
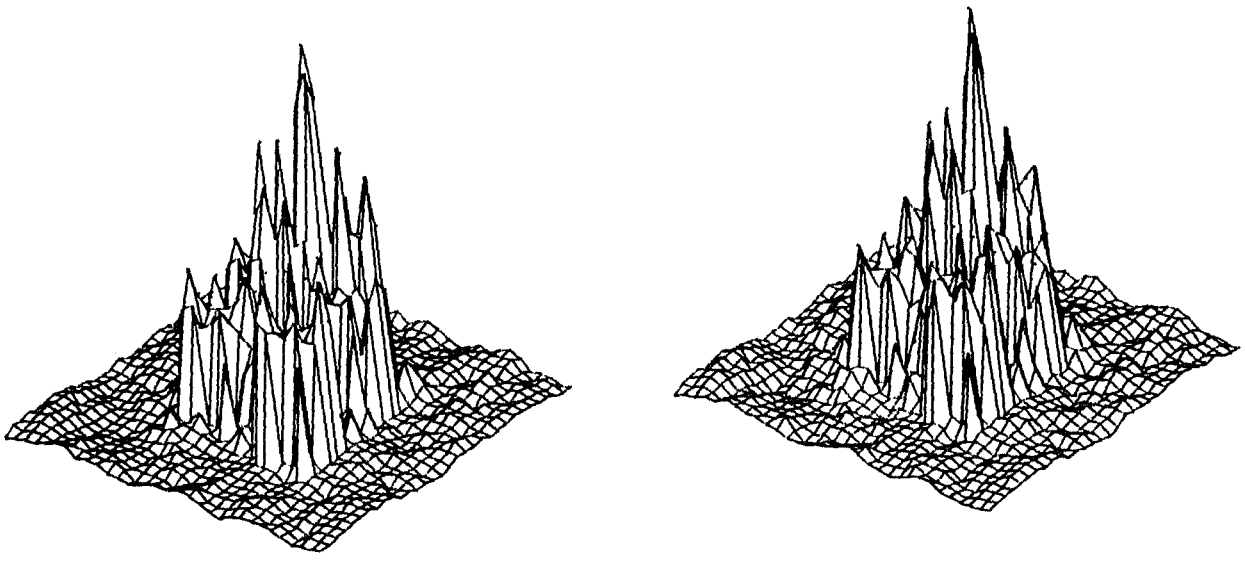


Figure 6.7: Error curves for reconstructions of Fig 6.1a from noisy visibility phase modulo π. In each subfigure (- - -) Support too large (——) Exact support (.....) Support too small (a) $N_F = 0.0$ (b) $N_F = 0.1$ (c) $N_F = 0.5$ (d) $N_F = 1.0$

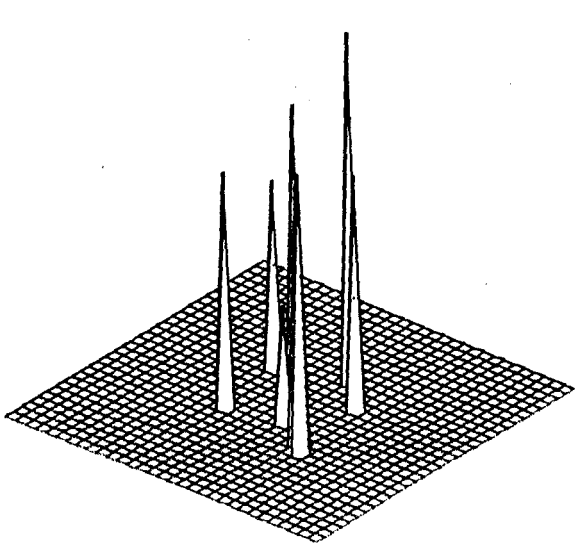


(a) (b)

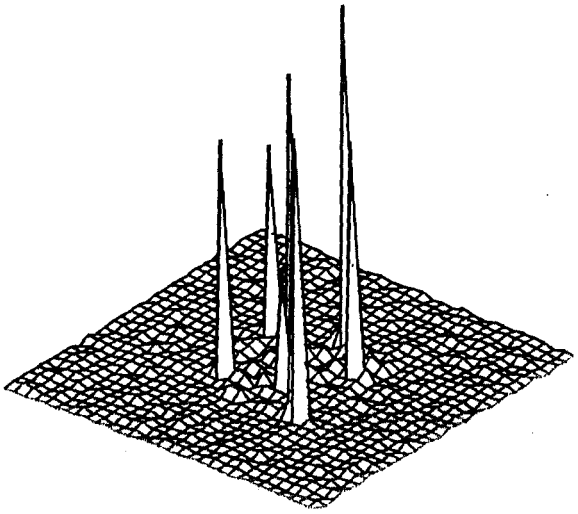


(c) (d)

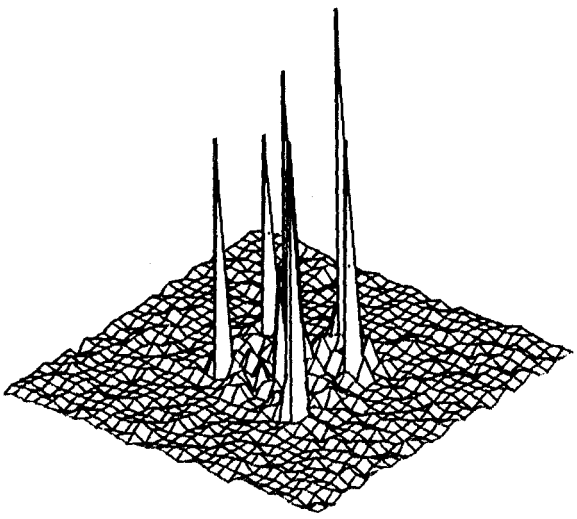
Figure 6.8: Reconstructions corresponding to the error curves in Fig 6.7 when the support is too large. (a) $N_F = 0.0$ (b) $N_F = 0.1$ (c) $N_F = 0.5$ (d) $N_F = 1.0$



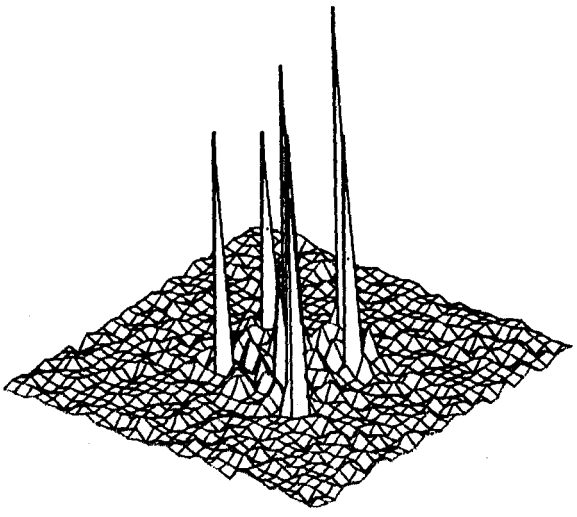
(e)



(f)

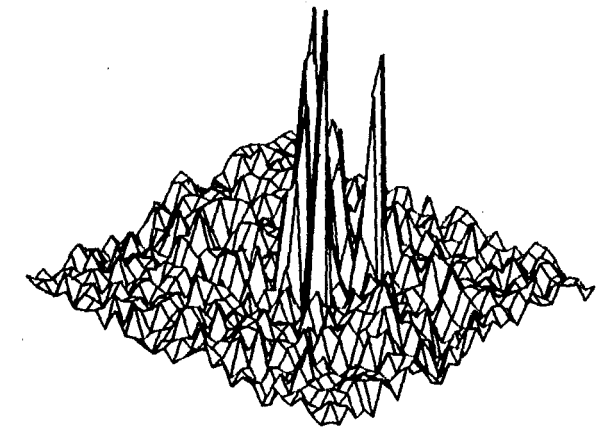


(g)

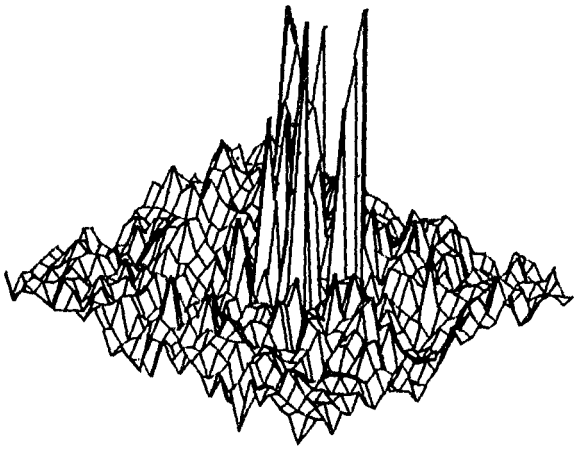


(h)

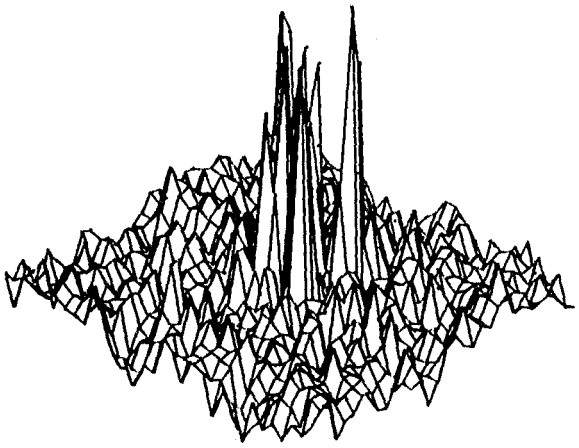
Figure 6.8: (continued) Reconstructions corresponding to the error curves in Fig 6.7 when the support is the correct size. (e) $N_F = 0.0$ (f) $N_F = 0.1$ (g) $N_F = 0.5$ (h) $N_F = 1.0$



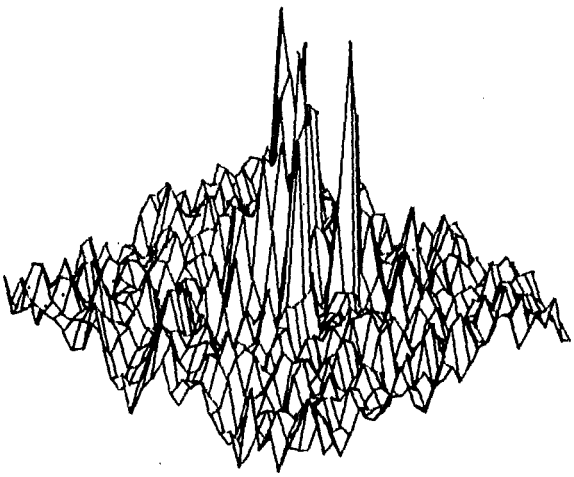
(i)



(j)

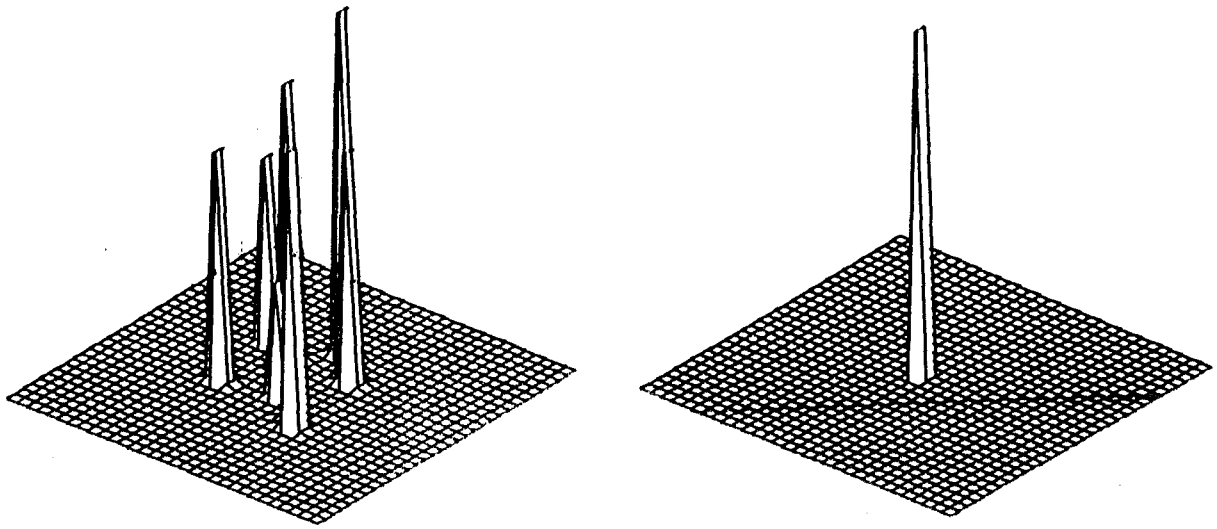


(k)



(l)

Figure 6.8: (continued) Reconstructions corresponding to the error curves in Fig 6.7 when the support is too small. (i) $N_F = 0.0$ (j) $N_F = 0.1$ (k) $N_F = 0.5$ (l) $N_F = 1.0$



(a) (b)

Figure 6.9: Reconstruction of Fig 6.1a when the Fourier phase is linearly shifted by an amount equivalent to a translation of less than a pixel spacing in image space (a) reconstructed image (b) computationally induced psf.

where ε_x is the sample spacing along the x-axis (see §1.5). Fig 6.9b shows the computationally induced psf in this case. How linear phase shifts corresponding to a translation of a fraction of a pixel arise is discussed in the following section, along with techniques for eliminating these unwanted phase shifts.

6.3 Blind deconvolution using the modified magnitude problem

The importance of the modified magnitude problem in blind deconvolution is emphasised initially in §2.5, where it is shown that provided that either the true image or the psf is symmetric it is possible to deduce the phase of the other component modulo π. In this section, the more general problem of deconvolution without constraining the image or the psf to be symmetric is addressed. It is convenient to assume that the convolution has two components $f(x,y)$ and $h(x,y)$, i.e.

$$g(x,y) = f(x,y) \odot h(x,y) \tag{6.13}$$

The extension to convolutions with more than two components is obvious and can be made without difficulty. The estimates made of $f(x,y)$ and $h(x,y)$ are denoted by $\hat{f}(x,y)$ and $\hat{h}(x,y)$ respectively.

Provided $g(x,y)$ is adequately sampled, arbitrarily closely spaced samples of $G(u,v)$ can be generated (as explained in §1.5) by zero packing $g(x,y)$ before computing

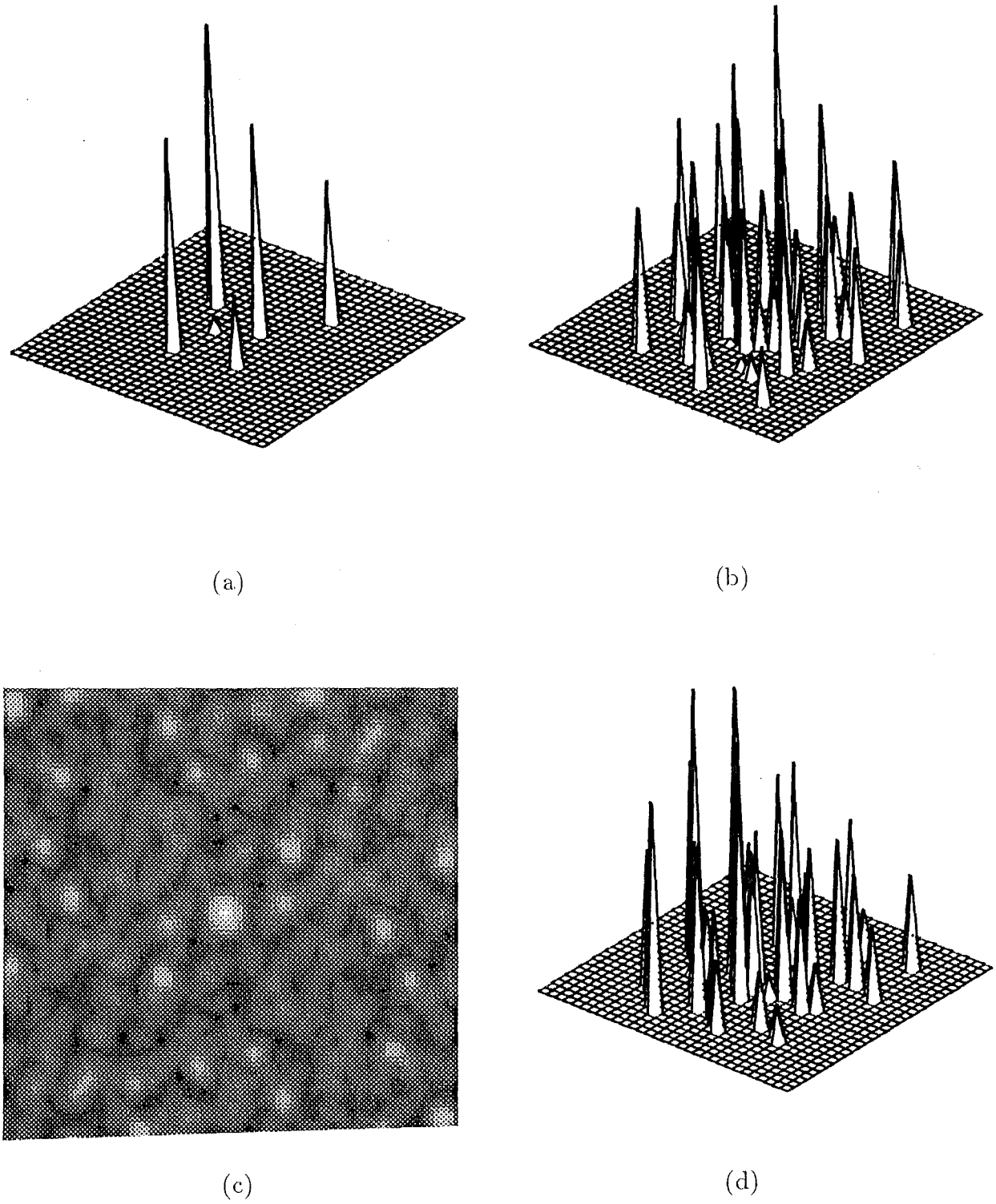


Figure 6.10: Relationship of $g(x,y)$ and $d(x,y)$. (a) psf $h(x,y)$ (b) $g(x,y)$ (c) $|G(u,v)|$ (d) $d(x,y)$.

the FFT. Discarding the phase of $G(u,v)$ then poses a Fourier phase problem. When several different starting points are tried it is apparent that some reconstructions correspond to the alternative image-form, $d(x,y)$ (as described in §4.3):

$$d(x,y) \longleftrightarrow D(u,v) = F(u,v)H^*(u,v) \tag{6.14}$$

Fig 6.10a shows a discrete pixellated psf $h(x,y)$ which is convolved with the $f(x,y)$ shown in Fig 6.1a to produce the convolution $g(x,y)$ shown in Fig 6.10b. The visibility magnitude $|G(u,v)|$ is shown in Fig 6.10c, whilst the alternative image-form $d(x,y)$ is shown in Fig 6.10d

The division of $G(u,v)$ by $D(u,v)$ yields

$$\frac{G(u,v)}{D(u,v)} = e^{i2\mathcal{P}[H(u,v)]} \tag{6.15}$$

whilst multiplying $G(u,v)$ and $D(u,v)$ gives

$$D(u,v)G(u,v) = |G(u,v)|^2 e^{i2\mathcal{P}[F(u,v)]} \tag{6.16}$$

Unfortunately (6.15) and (6.16) do not give the true phases of $F(u,v)$ and $H(u,v)$, for two reasons. Firstly, conversion from for example $2\mathcal{P}[F(u,v)]$ to $\mathcal{P}[F(u,v)]$, yields an ambiguity because

$$2\mathcal{P}[F(u,v)] \; mod \; 2\pi = 2\mathcal{P}[F(u,v) \pm \pi] \; mod \; 2\pi \tag{6.17}$$

Hence it is only possible to determine the phases of $F(u,v)$ and $H(u,v)$ modulo π. Secondly there arises a problem of linear phase shifts in image space, since it is only possible to recover the image-form of $d(x,y)$. For continuous (cf §1.5) positive images, phase shifts can be eliminated by translating $g(x,y)$ and $d(x,y)$ so that $B_d(x,y)$ and $B_g(x,y)$ are symmetric in image space. More care must be taken when dealing with discrete images since it is not always possible to centre a discrete image so that its support is exactly symmetric in image space. If for example $d(x,y)$ is misplaced by 1 pixel along the x-axis then the modulo π phases obtained correspond to

$$\mathcal{P}[F(u,v)]e^{i2\pi\frac{\varepsilon_x}{2}} \; mod \; \pi \tag{6.18}$$

and

$$\mathcal{P}[H(u,v)]e^{-i2\pi\frac{\varepsilon_x}{2}} \; mod \; \pi \tag{6.19}$$

and a fractional pixel shift in image space.

The elimination of unwanted phase shifts in the modulo π phases of the component visibilities can be cured by appropriate translation of $d(x,y)$. In order to do this it is necessary to consider a more general form of the discrete convolution than presented earlier in §4.2. Consider two square discrete images arbitrarily translated in image space,

$$f(x,y) = \sum_{m,n=-N_1}^{N_2} f(x - m\varepsilon, y - m\varepsilon) \tag{6.20}$$

and

$$h(x,y) = \sum_{m,n=-M_1}^{M_2} h(x - m\varepsilon, y - m\varepsilon) \tag{6.21}$$

where ε is the pixel spacing in image space. Thus the $B_f(x,y)$ and $B_h(x,y)$ are determined by N_1, N_2 and M_1, M_2 respectively.

The convolution of $f(x,y)$ and $h(x,y)$ is given by

$$g(x,y) = g(x-i\varepsilon, y-j\varepsilon) = \sum_{i,j=-(M_1+N_1)}^{N_2+M_2} \sum_{m,n=-M_1}^{M_2} f(x-m\varepsilon, y-n\varepsilon)h(x-(i-m)\varepsilon, y-(j-n)\varepsilon) \tag{6.22}$$

which has a support determined by $-(M_1+N_1)$ and (M_2+N_2). Because $B_g(x,y)$ is fixed by $-(M_1+N_1)$ and (M_2+N_2) it is possible to determine $B_g(x,y)$ from $B_f(x,y)$ and $B_h(x,y)$. Alternatively it is possible to determine $B_f(x,y)$ (or $B_h(x,y)$) from $B_g(x,y)$ and $B_h(x,y)$ (or $B_f(x,y)$).

To attempt to recover, for example $\mathcal{P}[F(u,v)]$ modulo π, without unwanted linear phase shifts one commences by choosing an initial $B_{\hat{f}}(x,y)$. Using $B_g(x,y)$ and $B_{\hat{f}}(x,y)$ it is then possible to determine the compatible $B_{\hat{h}}(x,y)$ as indicated in Fig 6.11a. $B_d(x,y)$ can then be determined from $B_{\hat{f}}(x,y)$ and $B_{\hat{h}}(-x,-y)$, as indicated in Fig 6.11b. The recovered image-form of $d(x,y)$ is then translated so that it lies within the support dictated by $S_d(x,y)$ before (6.16) is used to recover $\mathcal{P}[F(u,v)]$ modulo π.

It is important to understand why a blind deconvolution, which comprises two interconnected modified magnitude problems, is easier to solve than a single modified magnitude problem. It is of course essential to attack the modified magnitude problems jointly rather than separately. The reason is that there is more information contained in $g(x,y)$ and $d(x,y)$ than in the derived modulo π phases of $F(u,v)$ and $H(u,v)$. Because

$$G(u,v) = \hat{F}(u,v)\hat{H}(u,v) \tag{6.23}$$

then, if $\mathcal{P}[F(u,v)] = \mathcal{P}[F(u,v)]\ mod\ \pi$ it follows that $\mathcal{P}[H(u,v)] = \mathcal{P}[H(u,v)]\ mod\ \pi$ or, alternatively if $\mathcal{P}[F(u,v)] = (\mathcal{P}[F(u,v)]\ mod\ \pi)+\pi$ then $\mathcal{P}[H(u,v)] = (\mathcal{P}[H(u,v)]\ mod\ \pi)+\pi$. Furthermore $|G(u,v)| = |\hat{F}(u,v)\hat{H}(u,v)|$ which is not true when either $f(x,y)$ or $h(x,y)$ is blurred by a computationally induced psf.

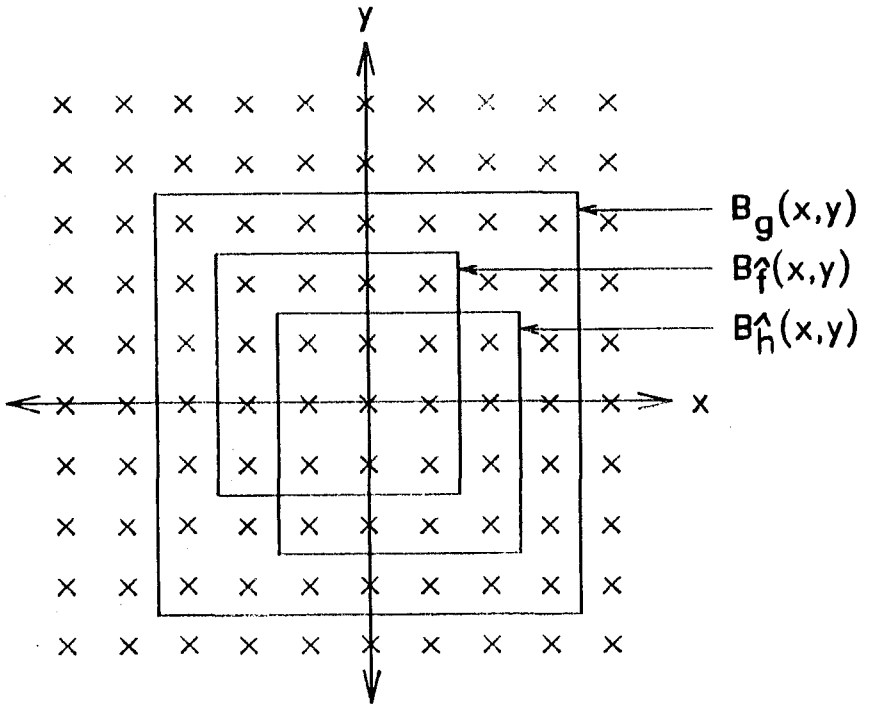
Now consider the deconvolution of the convolution shown in Fig 6.10b. Even when $h(x,y)$ has been recovered correctly the effects of an incorrectly estimating $B'_f(x,y)$ are readily apparent. Fig 6.12 shows

$$\mathcal{P}\left[\frac{\hat{F}(u,v)\hat{H}(u,v)}{G(u,v)}\right] \tag{6.24}$$

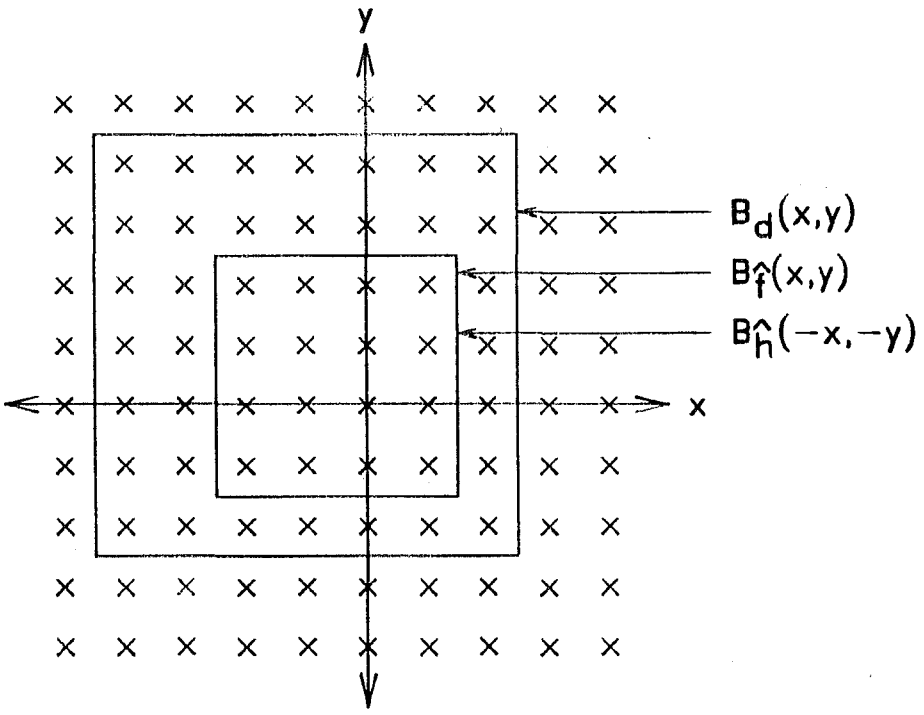
when $B_f(x,y)$ has been overestimated (Fig 6.12a) and underestimated (Fig 6.12b). Clearly there are large regions where (6.24) is not identically zero.

Similarly Fig 6.13 shows $|\hat{F}(u,v)\hat{H}(u,v)|$. Only when $B_{\hat{f}}(x,y) = B_f(x,y)$ is the product of the recovered magnitudes compatible with the magnitude of the convolution (Fig 6.10c). It should be remembered that, as described above, knowledge of two of $B_f(x,y)$, $B_h(x,y)$ and $B_g(x,y)$ unambiguously determines the remaining image-box. Hence in practice when $B_{\hat{f}}(x,y) \supset B_f(x,y)$ then $B_{\hat{h}}(x,y) \subset B_h(x,y)$ and vice versa.

There remains the problem of devising an algorithm which takes into account both the added phase and magnitude information inherent in the modified magnitude problems derived from a convolution. One approach is to attempt to recover both images simultaneously with different estimates of the components' supports. This allows the phases of the components to be constrained so that (6.24) is identically zero. Furthermore

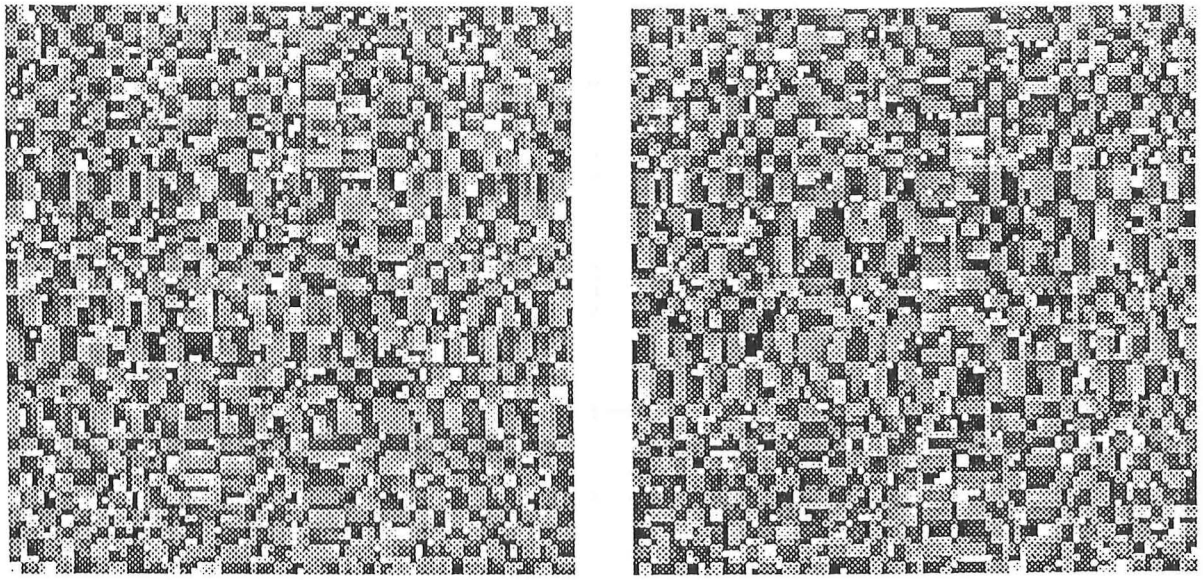


(a)



(b)

Figure 6.11: Determination of the support of $d(x,y)$. (a) Relationship of $B_{\hat{f}}(x,y)$ and $B_g(x,y)$ to $B_{\hat{h}}(x,y)$. (b) Relationship of $B_{\hat{f}}(x,y)$ and $B_{\hat{h}}(x,y)$ to $B_d(x,y)$.



(a) (b)

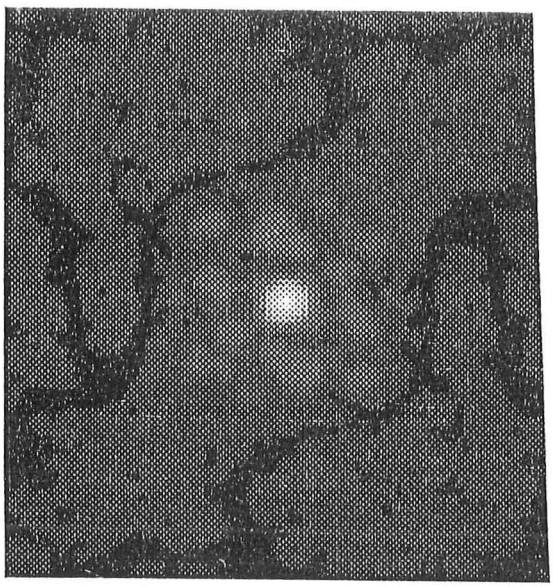
Figure 6.12: $\mathcal{P}\left[\frac{\hat{F}(u,v)\hat{H}(u,v)}{G(u,v)}\right]$ when (a) $B_{\hat{f}}(x,y) \supset B_f(x,y)$ (b) $B_{\hat{f}}(x,y) \subset B_f(x,y)$

it is possible to ensure that $|\hat{F}(u,v)\hat{H}(u,v)| = |G(u,v)|$. Although this algorithm can prove effective in many cases (Bates and Lane 1987a), further work has shown instances of poor convergence, especially when one or other of the images is symmetric (or nearly so). This is perhaps not surprising because the modulo π phase of any symmetric image is identically zero.

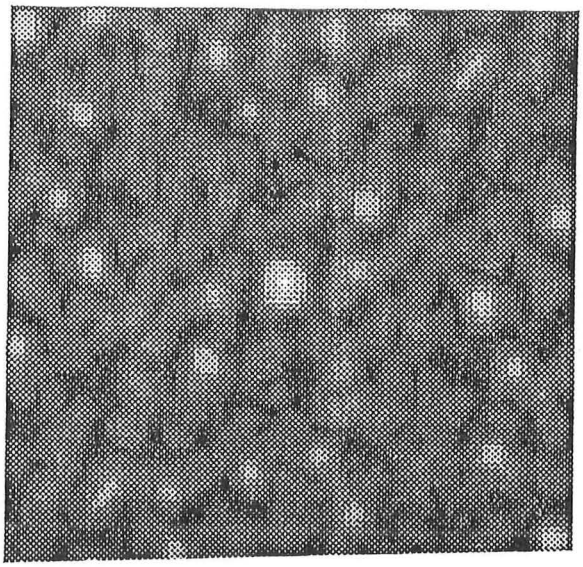
A better technique appears to be to recover one component from its modulo π phase for a number of different estimates of its support and then derive an estimate of the other component by Wiener filtering. This should always be possible except in the unlikely event that both components of the convolution are symmetric. If the recovered component, for example $\hat{f}(x,y)$, is blurred with a symmetric blurring function then $\hat{h}(x,y)$ is given by

$$\hat{h}(x,y) \longleftrightarrow \frac{H(u,v)}{S(u,v)} \tag{6.25}$$

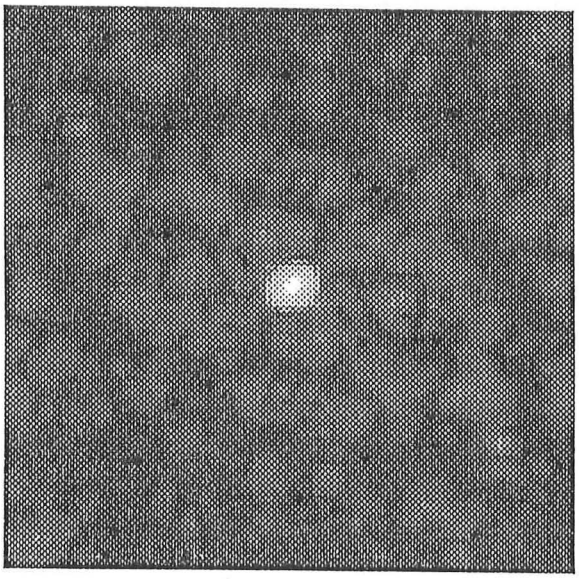
and is unlikely to be compact. Similarly, when $\hat{f}(x,y)$ is reconstructed within too small a support $\hat{h}(x,y)$ is also unlikely to be compact. Since $\hat{h}(x,y)$ is only compact when $\hat{f}(x,y)$ is equal to $f(x,y)$ the amount of $\hat{h}(x,y)$ within $B_{\hat{h}}(x,y)$ can be used as an error measure. ξ as defined in (1.29) can then be used to determine the correct values of $B_{\hat{f}}(x,y)$ and $B_{\hat{h}}(x,y)$. The advantage of this technique over comparing the convergence of E_I, as described in the previous section, is that it involves minimising a function and consequently should be more robust in the presence of noise. Fig 6.14 shows ξ for recovering $f(x,y)$ from the convolution shown in Fig 6.10b. The functional minimum at the correct support is readily apparent.



(a)



(b)



(c)

Figure 6.13: $|\hat{F}(u,v)\hat{H}(u,v)|$ when (a) $B_{\hat{f}}(x,y) \supset B_f(x,y)$ (b) $B_{\hat{f}}(x,y) = B_f(x,y)$ (c) $B_{\hat{f}}(x,y) \subset B_f(x,y)$

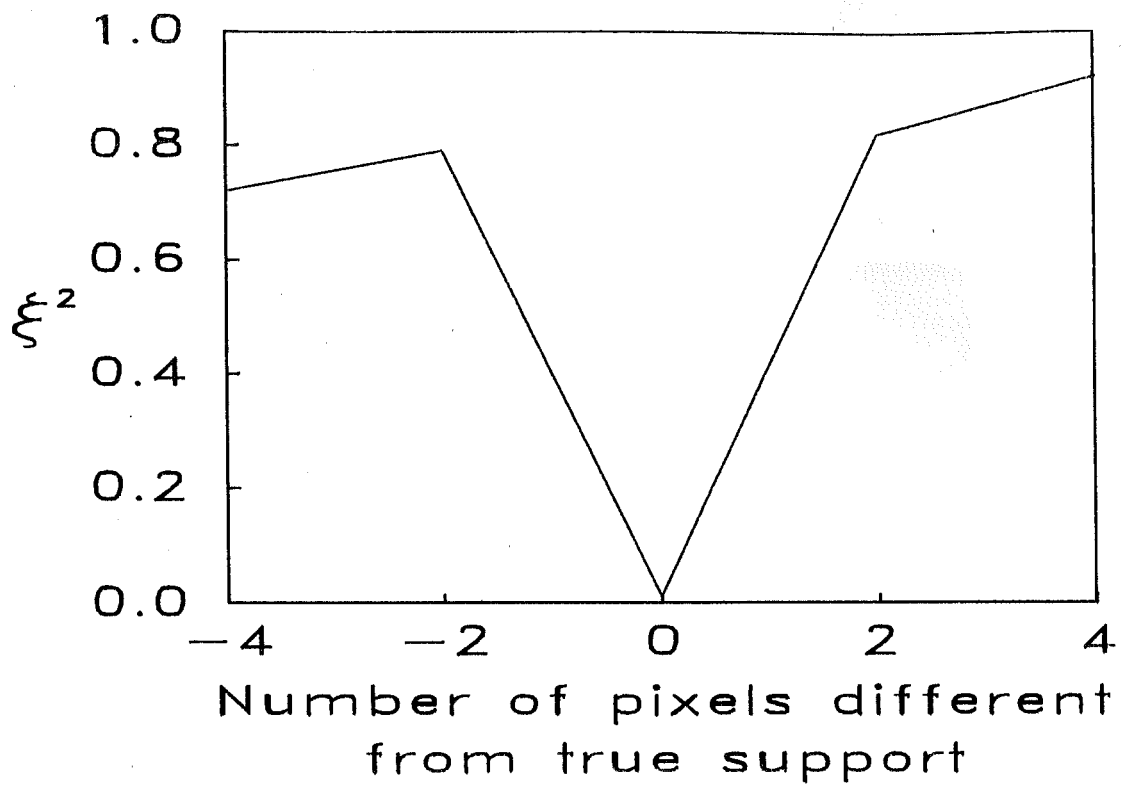


Figure 6.14: Compactness of $h(x,y)$ when $B_{\hat{f}}(x,y)$ differs from the true support $B_f(x,y)$.

6.4 Two-dimensional zero-and-add

Zero-and-add is a method of deconvolution applicable to an ensemble of differently blurred versions of a single image, such as occurs with a set of speckle images, i.e.

$$g_n(x,y) = f(x,y) \odot h_n(x,y) + c_n(x,y) \text{ for } n = 1,2,\ldots,N \tag{6.26}$$

or equivalently in Z-space

$$G_n(\zeta,\gamma) = F(\zeta,\gamma)H_n(\zeta,\gamma) + c_n(\zeta,\gamma) \text{ for } n = 1,2,\ldots,N \tag{6.27}$$

The basic principle relies on finding the zeros of each spectrum which are common to all members of the ensemble. In the absence of noise (i.e. $c_n(x,y) = 0$) these common zeros are the zeros of spectrum the true image,

$$\mathcal{Z}\{F(\zeta,\gamma)\} = \mathcal{Z}\{G_1(\zeta,\gamma)\} \cap \mathcal{Z}\{G_2(\zeta,\gamma)\} \cap \ldots \cap \mathcal{Z}\{G_N(\zeta,\gamma)\} \tag{6.28}$$

Zero-and-add was introduced for one-dimensional speckle images (Davey et al. 1987, Sinton et al. 1987, Sinton 1987). The purpose of this section is to illustrate how zero-and-add can successfully be implemented for two-dimensional images.

The example used is the 5 x 5 pixel object shown in Table 6.1. The data represents a non-square image embedded in a 5 x 5 pixel array. Although of small size many astronomical objects of interest are not much larger than the Airy disk of the telescope (§2.2), for example red giants in the largest available optical telescopes. In order to simulate blurring by a random phase field this object was convolved with an ensemble of

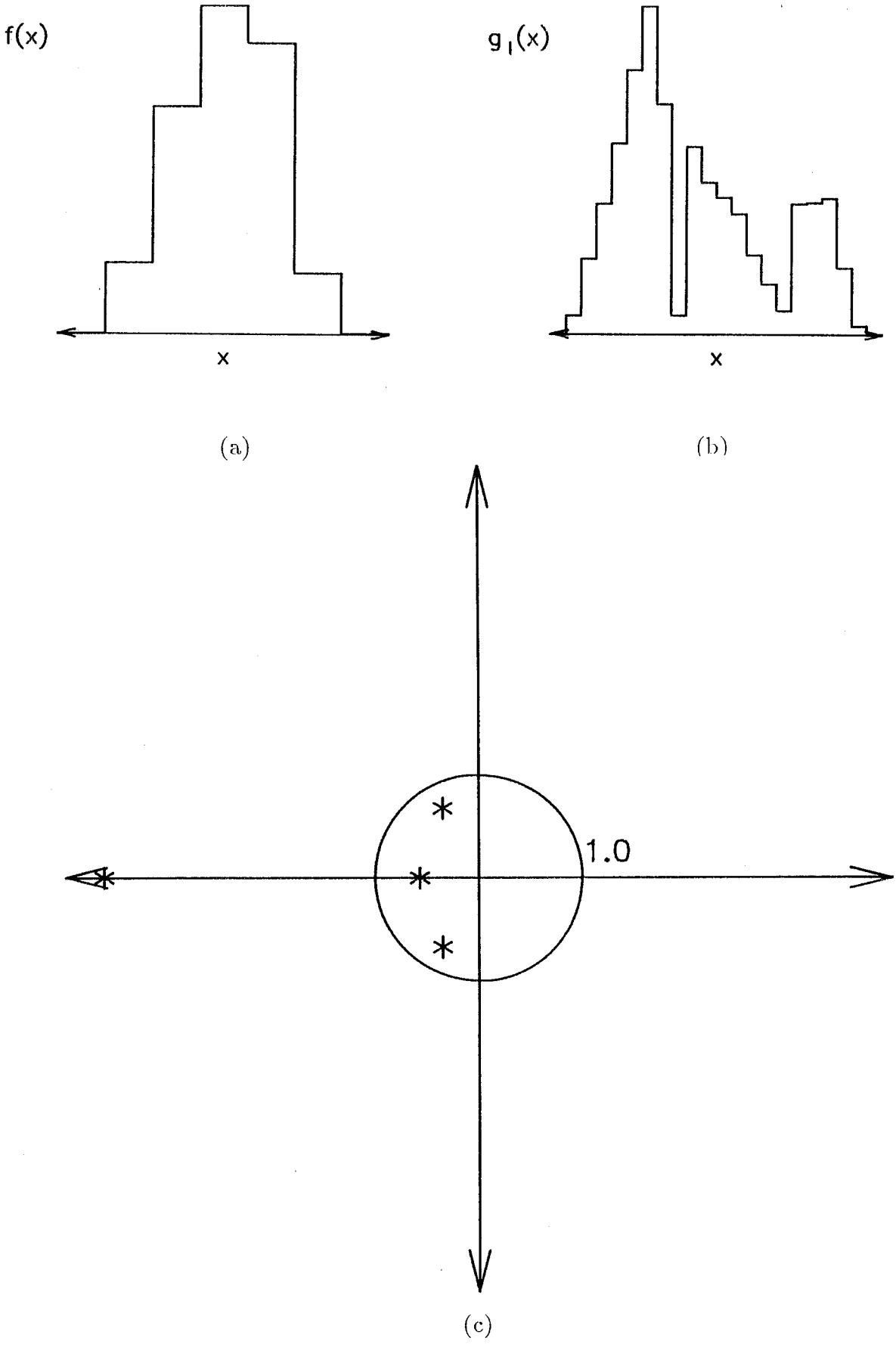


Figure 6.15: One-dimensional speckle images (a) the true image $f(x)$ (b) a typical speckle $g_1(x)$ (c) The zeros in the complex ζ-plane of the true image.

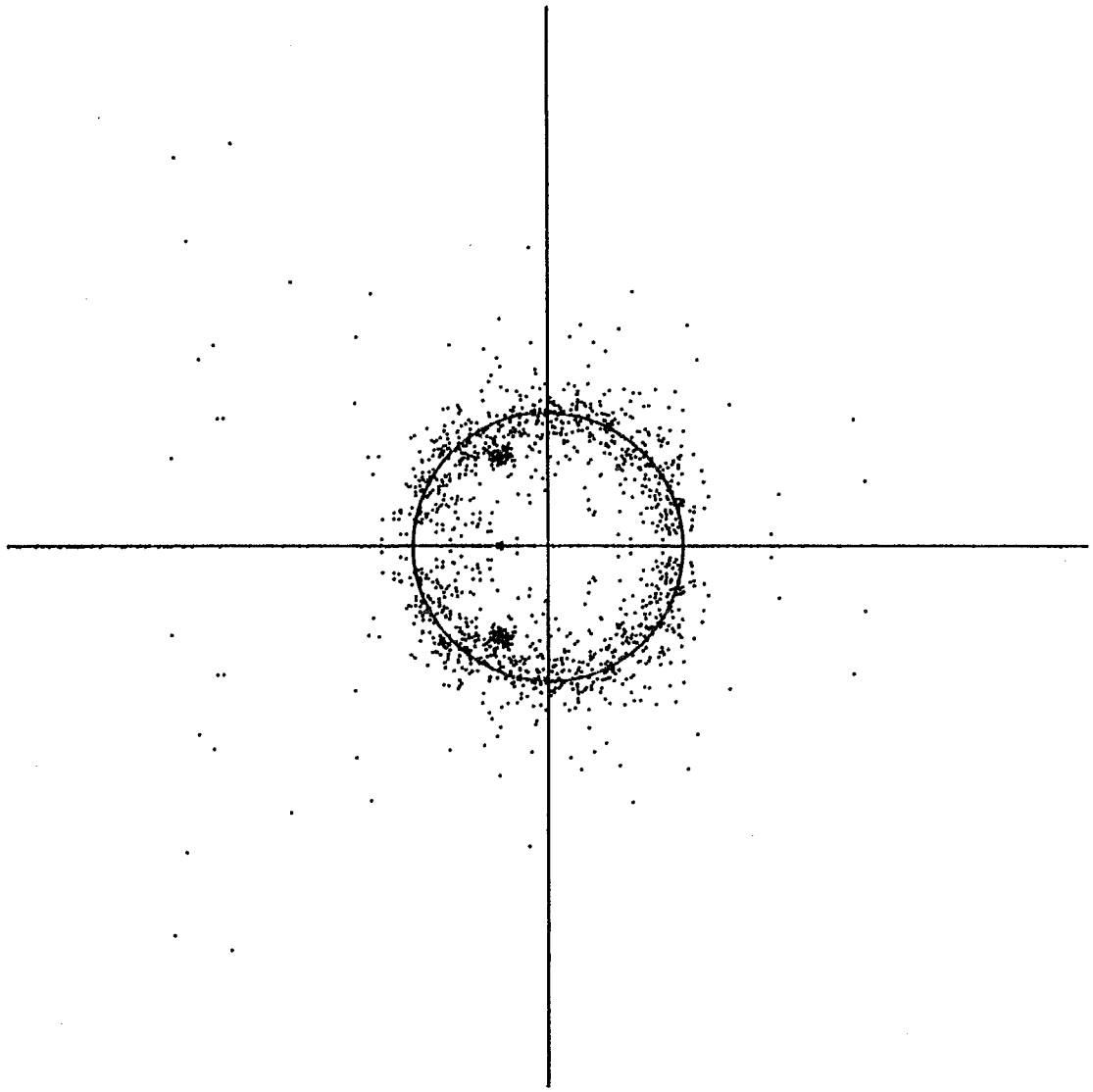


Figure 6.16: The superimposed zero-maps of all $\mathcal{G}_n(\zeta)$.

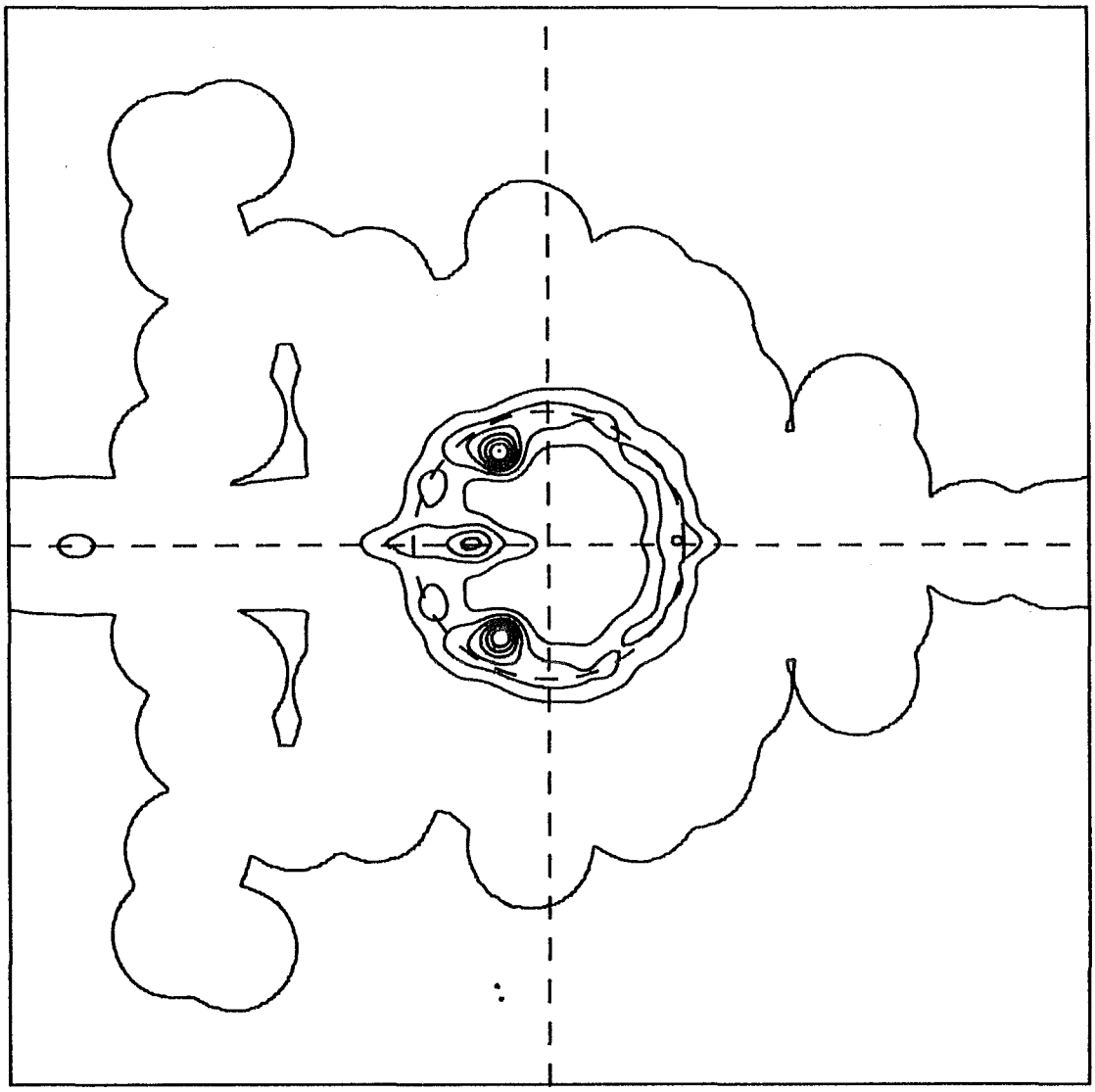


Figure 6.17: The zero density function.

$f_{m,n}$		m				
		0	1	2	3	4
n	0	0.00	0.00	0.10	0.13	0.00
	1	0.07	0.25	0.40	0.25	0.10
	2	0.13	0.30	0.20	0.30	0.07
	3	0.00	0.10	0.17	0.10	0.00
	4	0.00	0.07	0.07	0.05	0.00

Table 6.1: Values of $f_{m,n}$ for non-square image embedded in a 5 x 5 pixel array

16 x 16 random arrays of numbers. Noise was added to each convolution so that

$$\frac{< \int_{(x,y)} c(x,y)^2 dx dy >}{< \int_{(x,y)} g(x,y)^2 dx dy >} = 0.001 \tag{6.29}$$

where $< >$denotes the process of ensemble averaging. The simplest method of performing zero-and-add is to reduce each two-dimensional speckle image to a sequence of one-dimensional speckle images by the process of projection (§4.2). This is done in Z-space by setting either γ or ζ equal to a constant. The resulting ensemble of one-dimensional Z-transforms can then be used in a one-dimensional zero-and-add process. For example setting $\gamma = 1.0$, yields the following ensemble of one-dimensional convolutions

$$g_n(x) = f(x) \odot h_n(x) + c_n(x) \text{ for } n = 1, 2, \ldots, N \tag{6.30}$$

The one-dimensional true image $f(x)$ (corresponding to projecting the two-dimensional image of Table 6.1 with $\gamma = 1.0$), is shown in Fig 6.15a, along with its zeros in Z-space (Fig 6.15c). A typical $g_n(x)$ is shown in Fig 6.15b. Since $c(x)$ is not identically zero it is no longer possible to assume that the zeros of $F(\zeta)$ are a subset of the zeros of each $G_n(\zeta)$. In general, provided the noise is not too severe, the image zeros are randomly displaced only a small distance from their true positions. By contrast, the zeros due to the psf are uncorrelated between ensembles (Sinton 1986).

Consequently, when the zero maps of a large number of $G(u)$ are superimposed clusters of zeros form at the zero locations of the image, whilst the zeros of the psf contribute to a uniform background, Fig 6.16. It should be noted that zeros lying on the real axis tend to be displaced along the real axis rather than into the complex ζ-plane. Consequently, when simply plotting zero locations, zero clusters on the real axis are harder to discern than those with both real and imaginary components.

In order to convert this collection of discrete points into a continuous zero density map, each zero in Fig 6.16 is convolved with a gaussian. The resultant zero density function is shown in Fig 6.17. Although the peaks of the zero density map do indeed correspond to the locations of the original image zeros, these peaks are of different amplitudes, because not all of the original image zeros are equally sensitive to noise. In particular, zeros located well off the unit circle are particularly sensitive, a point noted by Sinton (1986). Hence in Fig 6.17 it is only possible to reliably determine three out of the original four image zeros because random fluctuations in the zero density exceed

$\hat{f}_{m,n}$		m				
		0	1	2	3	4
n	0	0.00	0.01	0.09	0.11	0.02
	1	0.06	0.22	0.40	0.29	0.09
	2	0.16	0.35	0.33	0.27	0.10
	3	0.07	0.16	0.20	0.19	0.04
	4	0.00	0.04	0.13	0.05	0.01

Table 6.2: Reconstruction of image given in Table 6.1 from 100 speckles

the height of the peak corresponding to the fourth zero. The loss of this zero effectively reduces the detail with which the true image is resolved.

In two-dimensions, however, it is possible to form as many zero-density maps as is desired, by using as many different values of ζ and γ as is deemed appropriate. When employing a linear equations approach to image recovery, as described in §4.8, there is no need to recover all the point zeros from each one-dimensional projection. Provided that the total number of zeros from all projections exceeds the the number of pixels in the image (cf §4.8), it is usually sufficient to recover just one or two zeros from each projection. The pixel values of the recovered image are listed in Table 6.2. Although E_T cannot be claimed to be small, the reconstruction replicates the shape of the image quite well.

The technique described in this section is still at a very early stage. There remains a significant amount of optimisation to be performed in both the one-dimensional zero-and-add process and the image recovery. As noted earlier in this section the positions of the true visibility zeros displaced far from the unit circle are very sensitive to noise. As a result, the peaks in the zero density map corresponding to these zeros are not well defined (Fig 6.17). This leads to two forms of error. Firstly spurious zeros may be located and secondly some of the true visibility's zero positions may be substantially misplaced.

Conventional least squares, as used in the above example, is very sensitive to incorrect data. Hence a weighted least squares technique (Taylor 1982) should result in significantly better reconstructions, provided a reasonable estimate of the error in zero positions can be made. Further possible improvements to the technique outlined in this section are suggested in chapter 7.

Chapter 7

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

There are many possiblities for continuing the work discussed in this thesis, both in blind deconvolution and phase retrieval. One fundamental issue worth closer examination is how to represent a compact image. Although sampling has proved adequate in most situations, it appears that discrete prolate spheroidal wavefunctions may be capable of providing the basis for a new generation of algorithms.

Another area warranting further study is the question of making more efficient use of phase closure in two-dimensional phase retrieval, which is the subject of §7.2. In my opinion an algorithm which enforces approximate phase closure over the entire (u, v)-plane is likely to lead to efficient phase retrieval, especially when dealing with noisy Fourier magnitude data.

Means of improving the convergence of iterative phase retrieval algorithms are needed. This is the subject of §7.3, where the major emphasis is on finding a good starting estimate. At the very least a good starting estimate accelerates the convergence of an iterative algorithm (Won et al. 1985), whilst in situations where the constraints are not particularly strong a good starting estimate can mean the difference between an acceptable and meaningless reconstruction.

The final sections of this chapter relate to possible improvements to the new methods of deconvolution introduced in this thesis. §7.4 deals with deconvolution based on separating the zero-sheets corresponding to the components of the convolution, whilst §7.5 discusses possible improvements to blind deconvolution based on recovering the modulo π phase of the components. The last section (§7.6) deals with two-dimensional zero-and-add. Because the method described in §6.4 does not take full advantage of the analytic properties of two-dimensional images, it appears likely that considerable improvements to the results presented in §6.4 should be possible.

7.1 Prolate spheroidal wavefunctions

All practical images are of finite energy and as a consequence they must be effectively compact in both Fourier and image space. Although it is possible to represent a compact image by samples in either Fourier or image space, this representation is not optimal (§1.6,

Landau and Pollack 1962). To show this, consider a one-dimensional image $f(x)$ which is effectively compact (§1.6) in image space within the interval $\left[-\frac{X}{2}, \frac{X}{2}\right]$ and in Fourier space within the interval $[-W, W]$. As noted by Slepian (1983), $f(x)$ can be represented by the weighted sum of $(2WX + 1)$ prolate spheroidal wavefunctions (PSWFs) whose functional forms are determined solely by the product WX.

When representing a compact image by samples, however, it is necessary to use more than $(2WX + 1)$ samples to model a compact image, especially when the product WX is small. The drawback to this approach is that only a subset of the images representable by the increased number of samples are in fact consistent with the constraint of compactness. Because the dimension of the space spanned by a sampled representation is larger than the dimension of the actual space, it is correspondingly more difficult to find a feasible solution (Trussell and Civanlar 1984). Also, using an increased number of samples allows the image's Fourier transform to exist in a wider space $[-W', W']$ where $W' > W$. Consequently it is necessary to alternate between image and Fourier space to ensure that the constraint of compactness in Fourier space is being met.

Discrete PSWFs have major advantages over continuous PSWFs for representing a compact image, since they are much easier to compute. Slepian (1978) describes how the discrete PSWFs can be calculated by solving a finite matrix eigenvalue problem, a familiar problem in many branches of engineering.

As an example of how discrete PSWFs could be used as an alternative to sampling functions, consider the Fourier phase problem. Firstly, the discrete PSWFs are calculated from the product WX. Then, assuming the image is exactly compact, all that remains is to fit the first $(2WX + 1)$ discrete PSWFs to the Fourier magnitude. Since the discrete PSWFs are compact in image space by definition, there is no longer the need to iterate between Fourier and image space (as required when using sampling functions). Hence the problem reduces to optimally fitting a set of basis functions to the available data.

7.2 Two-dimensional phase closure

§4.2 introduces a technique for phase retrieval based on enforcing phase closure. To briefly recap the technique, a closed circuit in the (u, v)-plane is formed, the boundaries of which are straight lines on which the Fourier transforms of projections in image space are defined. Referring to Fig 4.7 and (4.15) it is apparent that the total phase around this circuit must equal zero modulo 2π. The algorithm as it stands performs zero-flipping to find the phase distributions along the one-dimensional boundaries which yield exact phase closure.

The practical difficulties presented by this crude algorithm, even in the absence of noise, are immense. §4.2 notes that the computation required is an exponential function of the image size. Furthermore, in the presence of noise it is no longer possible to rely on exact phase closure around a circuit in Fourier space. In my opinion, considerable improvements to the algorithm are possible by utilising the results of §3.4, where it is noted that it is usually acceptable to model an image with a finite number of zeros.

Consider the typical zero distribution for a one-dimensional image shown in Fig 7.1. It is a property of the the Fourier transform of an object of compact support that its zeros tend to the real axis as $\mathcal{R}[u] \to 0$, as is explained in detail by Requicha (1980). Intuitively, because all compact real world images are of finite energy, $F(u) \to 0$ as

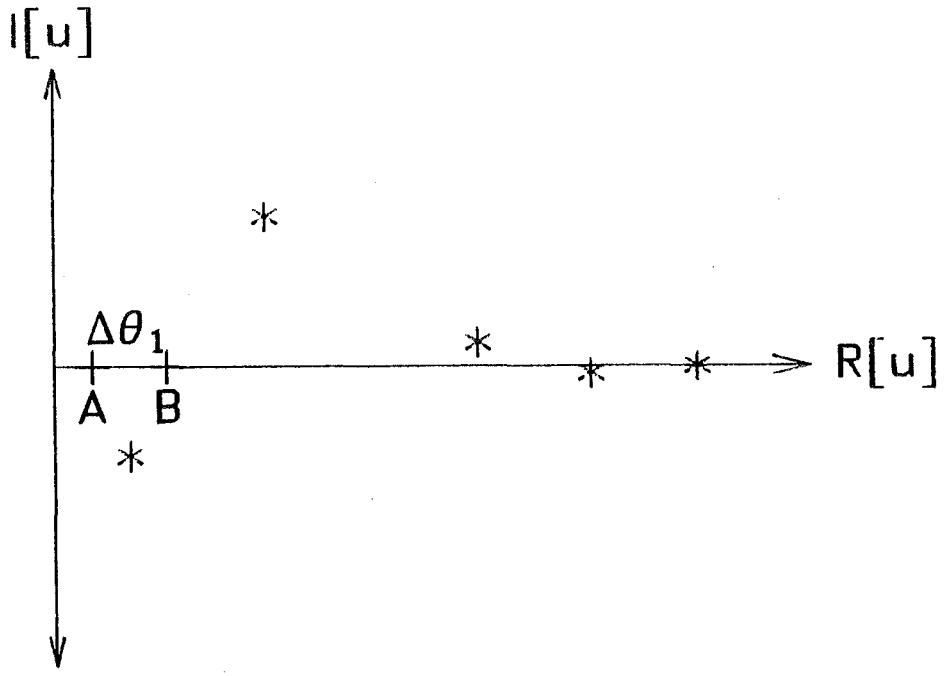


Figure 7.1: Typical zero distribution in Fourier space for a one-dimensional compact image.

$u \to \infty$. Thus it is not unreasonable to expect the zeros of $F(u)$ to tend towards the real axis when $u \to \infty$.

Now consider the phase difference $\Delta\theta_1$ between points A and B in Fig 7.1. The contribution made by each zero is proportional to the distance of that zero from points A and B. As a result the low frequency phase is usually determined by a few low frequency zeros (cf §3.4). Since the low frequency phase in turn determines the general structure of an image, although there may be a large number of image-forms, in general there are very much fewer general shapes for the image-forms (a point noted in practical investigations by both Fright (1987) and Fienup (1978)).

It should be possible to use approximate phase closure, using only small circuits in Fourier space, and only considering the zeros which are close to each circuit. It should then be possible to combine the information obtained from several circuits to uniquely determine the true phase. Because this technique would only involve a small number of zeros for any one calculation there should no longer be an exponential rise in computation with image size. Furthermore, since it would no longer be necessary to employ exact phase closure, the technique could be made quite robust in the presence of noise.

7.3 Accelerating phase retrieval by optimal choice of starting estimate

The importance of a good starting estimate is well undestood in iterative processing. At the very least a good starting estimate accelerates convergence of an iterative algorithm. In many cases where the constraints are not particularly strong, a good starting estimate can mean the difference between an acceptable and meaningless reconstruction.

One possible area for future research would be to apply a similar approach to the phase problem as it arises in speckle imaging applications. Assuming the current estimate of the true image has the correct Fourier magnitude it is possible to model it as the convolution of the true image and a blurring function. Invoking the notation of §5.1, the current estimate of the true visibility can be written as

$$K(u,v) = F(u,v)e^{i\Psi(u,v)} \tag{7.1}$$

where $\Psi(u,v)$ is a real function of u and v.

(7.1) bears a marked similarity to the situation in speckle imaging where the image's visibility phase is effectively randomised. Hence, by applying Fienup's iterative algorithms to a number of different starting images it should be possible to form an ensemble of images which could be used in, for example, a variation of SAA (cf Robinson and Bates (1982)). The resultant image could then be used as the initial estimate for further Fienup processing. A technique of this form could, to a large degree, overcome the stagnation resulting from a poor choice of starting image (§5.7).

Another method which warrants further investigation is optimally positioning the estimated support. A commonly occurring problem is that the recovered image is translated relative to the assumed support (§5.7). A possible way of averting this difficulty is to periodically shift the image-box so as to maximise the enrgy within the support (Fright 1984; Fienup and Wackerman 1986).

I have made a preliminary investigation of a variation of this approach. Rather than move the support to maximise the energy within the support, the image is translated a few pixels in every direction and then each image is used as the starting image for a small number of iterations. The output image with the smallest E_I is then used as a new basis and the process is repeated. After a few of these cycles, a conventional Fienup iteration is employed. This technique successfully averts stagnation due to the image-form attempting to reconstruct in a position translated relative to the estimated support. Although there is a marked improvement in the convergence of the iterative loop when it is implemented in this manner, it is not clear whether the extra initial computation is justifiable.

7.4 Zero based blind deconvolution

The technique for zero-based deconvolution introduced in §4.7 can, in the absence of noise, deconvolve quite large objects. As noted in §4.9, the addition of noise to the convolution causes the zero-sheets corresponding to the components of the convolution to become linked by bridges. Since these bridges exist in a four-dimensional space, their visualisation poses severe difficulties which are only partly resolved by the techniques introduced in §4.6.

In order to develop an algorithm which can robustly deconvolve noisy convolutions, it will be necessary to develop a deeper understanding of the behaviour of zero-sheets. In particular, some means must be found to relate the zero-sheets of aliased convolutions with the zero-sheets of convolutions sampled at the Nyquist frequency. Ideally, the algorithm should be capable of starting with a low frequency version of the zero-sheet, afterwards slowly building up the spatial frequency coverage as the resolution of the image is increased. Thus, rather than deal with the full zero-sheet of the visibility

of, for example a 63 x 63 pixel image, it should be possible to reduce the problem to a 9 x 9 pixel image which models only the low frequency portion of the Fourier spectrum.

Since, as noted in §4.9, the number of bridges is related to the number of pixels in the image, reducing the effective size of the convolution would considerably reduce the computation required. It should also be far simpler than isolating and correcting the bridging between the zero-sheets of the components of a noisy convolution.

7.5 Phase-based blind deconvolution

§6.3 deals with an iterative method of blind deconvolution. Utilising the notation of §6.3, the deconvolution procedure can be divided into two stages. Firstly, the convolution's visibility magnitude is used to pose a Fourier phase problem which is solved using techniques introduced in chapter 5. Depending on the initial choice of starting point, this yields either

$$g(x,y) = f(x,y) \odot h(x,y) \tag{7.2}$$

or

$$d(x,y) = f(x,y) \odot h^*(-x,-y). \tag{7.3}$$

On using (7.2) and (7.3), it is possible to derive $\mathcal{P}[F(u,v)]$ and $\mathcal{P}[H(u,v)]$ both modulo π. The second stage of the deconvolution consists of recovering $f(x,y)$ and $h(x,y)$ from their respective phases modulo π. The major difficulty with this second stage is correctly estimating $B_f(x,y)$ and $B_h(x,y)$. It appears likely, however, that refinements of the final iterative procedure outlined in §6.3 should be capable of overcoming these problems.

The first stage poses a number of difficulties when dealing with more detailed images, for example Fig 7.2. This is because it is very difficult to recover $d(x,y)$ precisely. As can be seen from Fig 7.2c, the convolution of a pair of positive objects exhibiting pronounced detail tends to be a relatively featureless image exhibiting little apparent detail. Recovery of "foggy" image-forms poses considerable difficulties, which are detailed in Fright (1984).

Fig 7.3 shows

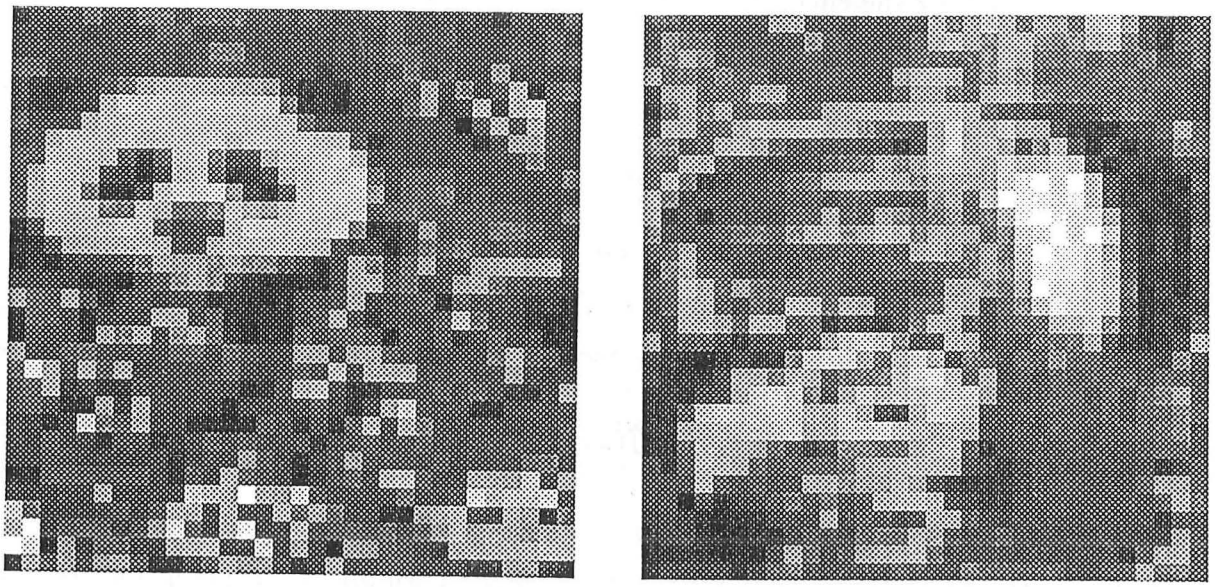
$$\mathcal{P}\left[\frac{\hat{D}(u,v)}{D(u,v)}\right] \tag{7.4}$$

where $\hat{D}$ is the estimated visibility obtained using Fienup's hybrid input-output algorithm, which is described in chapter 5. Although phases close to the origin of Fourier space have been estimated accurately, the same can not be said for the phases further out. Because it is impossible to recover the Fourier magnitude correctly in regions where the Fourier phase is incorrect, the reconstructed images shown in Fig 7.4 are equivalent to lowpass filtered versions of the images in Fig 7.2.

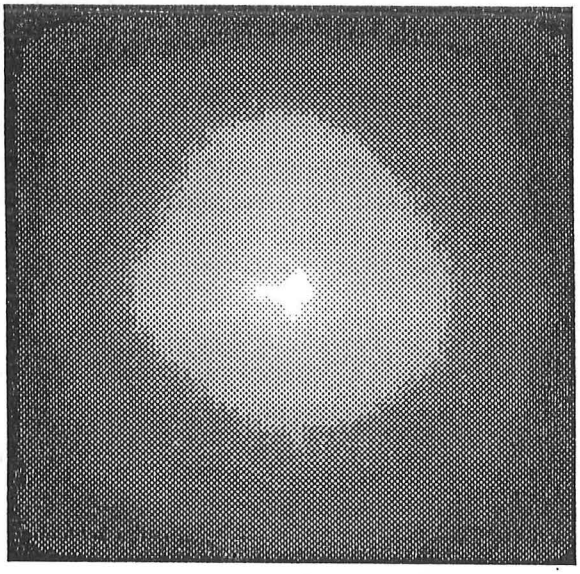
Fright (1984) details a number of procedures for dealing with "foggy" images. It is likely that appropriate adaption of these techniques could enhance the quality of the reconstructions shown in Fig 7.4.

7.6 Two-dimensional zero-and-add

The technique for two-dimensional zero-and-add, described in §6.4, relies on reducing the two-dimensional speckle images to a set of ensembles of one-dimensional



(a) (b)



(c)

Figure 7.2: Example of the convolution of two 32 x 32 pixel positive images. (a) first 32 x 32 pixel image (b) second 32 x 32 pixel image (c) the 63 x 63 pixel convolution of the images shown in Fig 7.2a and 7.2b

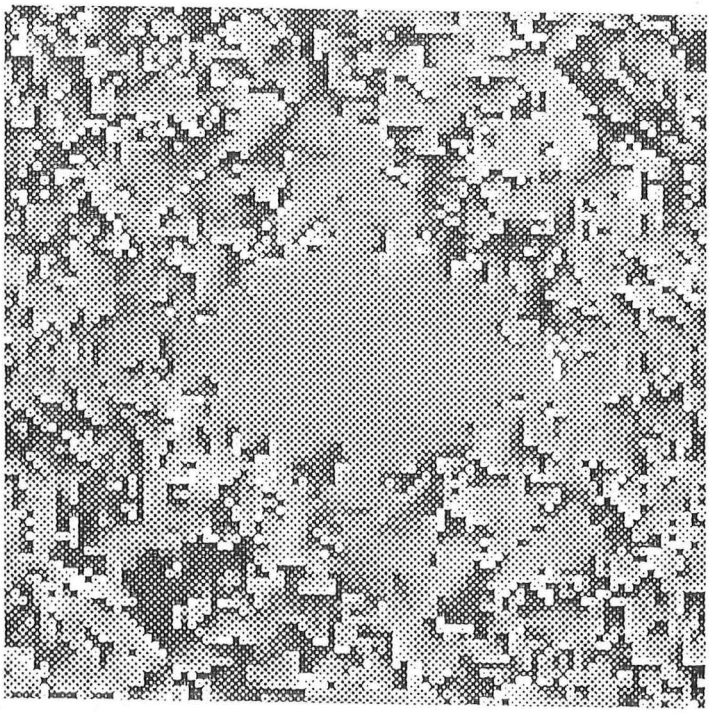
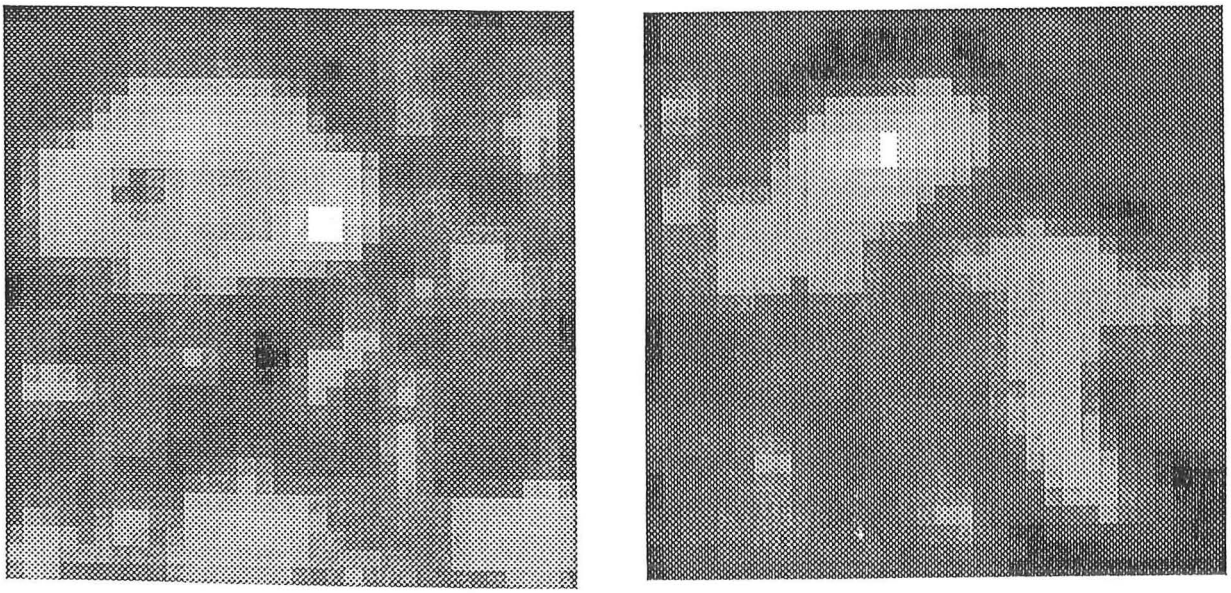


Figure 7.3: Phase of the ratio of $\mathcal{P}[\hat{D}(u,v)]$ and $\mathcal{P}[D(u,v)]$ (cf (7.3)). Quantised from $-\pi$ (black) to π (white).



(a) (b)

Figure 7.4: Reconstructions, of the two images shown in Fig 7.2a and b, formed by deconvolving Fig 7.2c

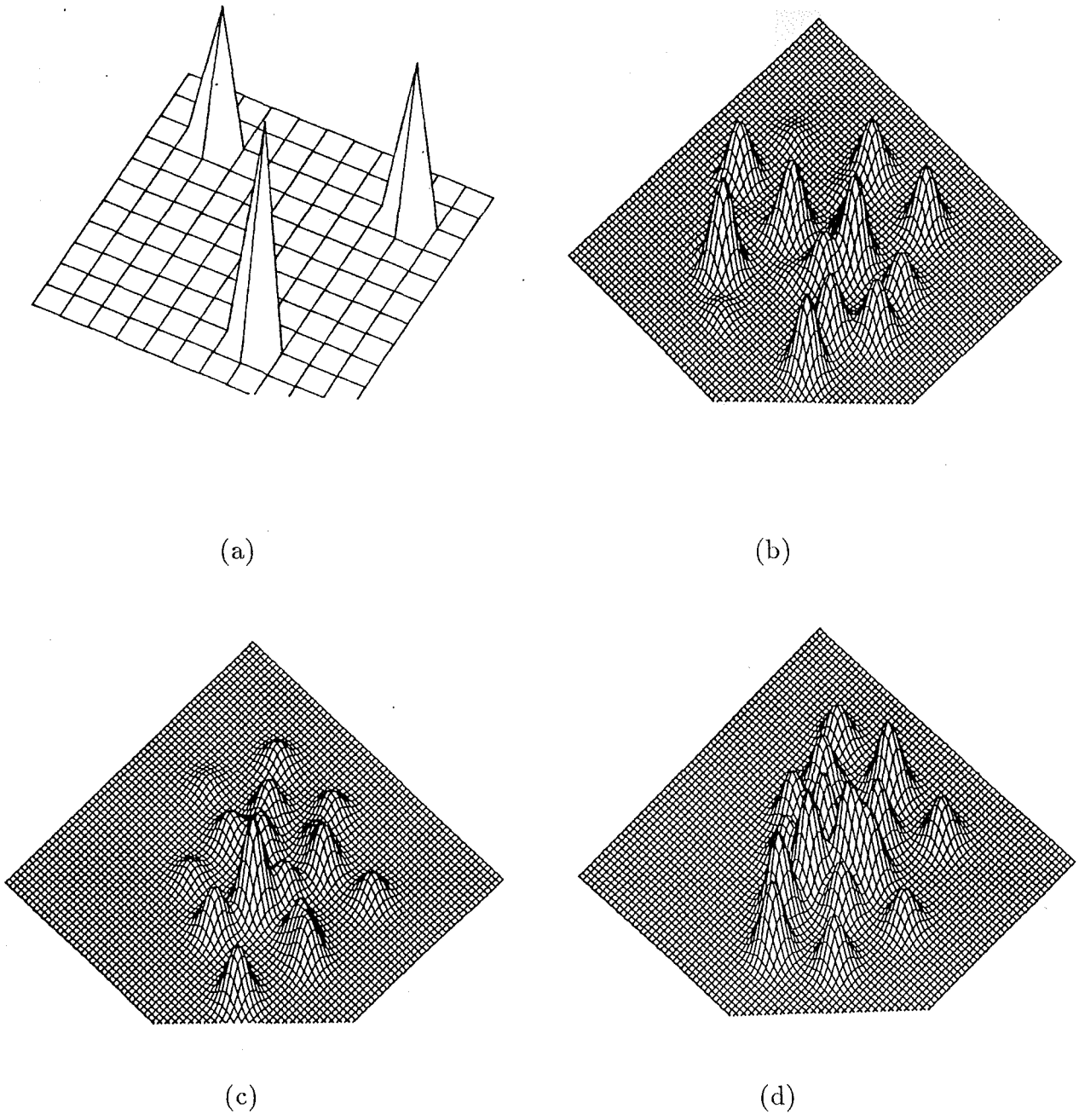


Figure 7.5: Simulated speckle images. (a) the true image (b) Speckle #1 (c) Speckle #2 (d) Speckle #3

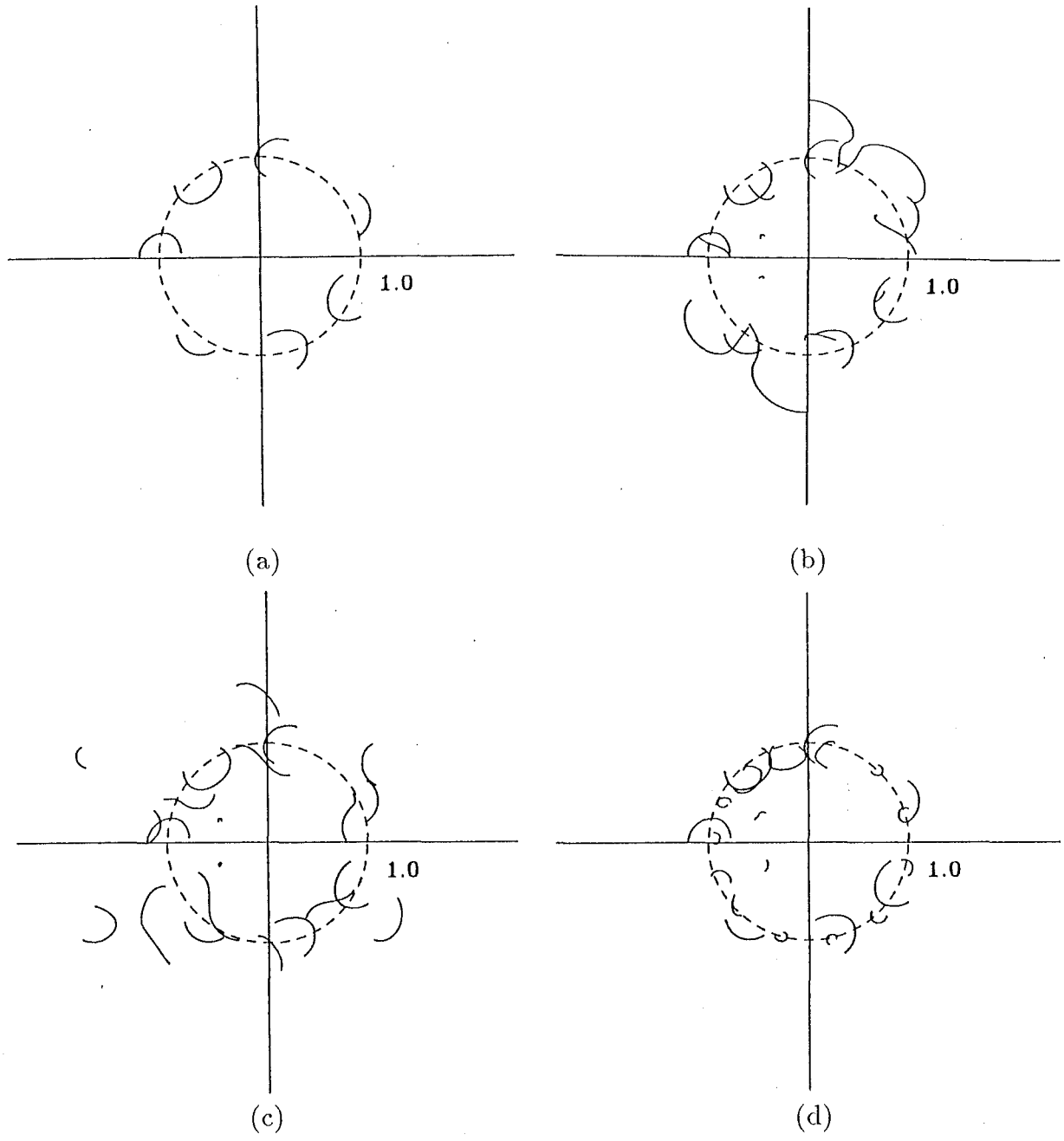


Figure 7.6: Sections of zero-contours corresponding to $\rho = 1.0$, $\eta = [0, 0.42]$ (a) true image (b) Speckle #1 (c) Speckle #2 (d) Speckle #3

speckle images. A considerable amount of information is discarded with this approach because a zero-sheet, rather than being a disjoint collection of points, is in fact a single continuous surface. It should be far more efficient to compare segments of zero-contours or areas of zero-sheets to see if they are common to each speckle image.

A preliminary illustration of the above suggestion is shown in Figs 7.5 and 7.6. Fig 7.5a shows a point image, whilst Figs 7.5b-d show three simulated speckle images. Employing the notation introduced in chapter 4, the zero-contours shown in Fig 7.6 correspond to fixing $\rho = 1.0$ and varying η from 0 to 0.42. It is apparent that the zero-contours of the true image are present in those of each of the speckle images. The extension to zero-and-add introduced in §6.4 relies on comparing zero maps formed by plotting points, sampled from the two-dimensional zero-sheet. Performing a similar process except with segments of zero-contours, or indeed sections of a zero-sheet should be inherently more robust. Intuitively it would appear far less likely for an analytic set of points to coincide randomly than it would be for a single point. As a result, those components of the zero-sheets which are in fact common to each speckle image's visibility should be easier to identify.

Finally, it would be desirable to analyse the effects of noise on two-dimensional zero-sheets in more detail. Clearly, not all sections of the zero-sheet are equally sensitive to noise, and this should be taken into account when recovering the image-form. Parts of the zero-sheet which are likely to have been substantially affected by noise must be given appropriately less weighting than other parts, if the image-form is to be recovered robustly.

References

[Ables 1974] J. G. Ables. Maximum entropy spectral analysis. *Astronomy and Astrophysics Supplement Series*, 15:383–393, 1974.

[Akaike 1974] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974.

[Anderson and Anderson 1986] R. C. Anderson and C. S. Anderson. Signal processing using only fourier phase. *Optical Engineering*, 25:1316–1319, 1986.

[Andrews and Hunt 1977] H. C. Andrews and B. R. Hunt. *Digital Image Restoration.* Prentice-Hall, New Jersey, 1977.

[Arsenault and Chalasinska-Macukow 1983] H. H. Arsenault and K. Chalasinska-Macukow. The solution to the phase retrieval problem using the sampling theorem. *Optics Communications*, 47:380–386, 1983.

[Astrom 1974] K. J. Astrom. Maximum likelihood and prediction error methods. *Automatica*, 16:551–574, 1974.

[Barakat and Newsam 1984] R. Barakat and G. Newsam. Necessary conditions for a unique solution to two-dimensional phase recovery. *Journal Mathematical Physics*, 25:3190–3193, 1984.

[Bates and Cady 1980] R. H. T. Bates and F. M. Cady. Towards true imaging by wideband speckle interferometry. *Optics Communications*, 32:365–369, 1980.

[Bates and Lane 1987a] R. H. T. Bates and R. G. Lane. Automatic deconvolution and phase retrieval. In *Proceedings SPIE*, 1987.

[Bates and Lane 1987b] R. H. T. Bates and R. G. Lane. Automatic deconvolution and phase retrieval. In *Proceedings of the joint workshop on high-resolution imaging from the ground using interferometric techniques, Oracle, Arizona*, pages 71–73, January 1987.

[Bates and Lane 1987c] R. H. T. Bates and R. G. Lane. Deblurring should now be automatic. In *6th Pfefferkorn Conference on Image and Signal Processing in Electron Microscopy, Niagara Falls, Ontario, Canada*, April 1987.

[Bates and Mnyama 1986] R. H. T. Bates and D. Mnyama. The status of practical Fourier phase retrieval. In P. W. Hawkes, editor, *Advances in Electronics and electron physics*, pages 1–64, Academic Press, 1986.

[Bates and McDonnell 1986] R. H. T. Bates and M. J. McDonnell. *Image Restoration and Reconstruction.* Clarendon Press, Oxford, 1986.

[Bates and Robinson 1982] R. H. T. Bates and B. S. Robinson. A stochastical imaging procedure. *Acoustical Imaging*, 12:185–191, 1982.

[Bates and Tan 1985] R. H. T. Bates and D. G. H. Tan. Fourier phase retrieval when the image is complex. In R. H. T. Bates and A. J. Devaney, editors, *Proceedings SPIE, Inverse Optics II*, pages 54–59, 1985.

[Bates 1969] R. H. T. Bates. Contributions to the theory of intensity interferometry. *Monthly Notices of the Royal Astronomical Society*, 142:413–428, 1969.

[Bates 1978a] R. H. T. Bates. Fringe visibility intensities may uniquely define brightness distributions. *Astronomy and Astrophysics*, 70:L27–L29, 1978.

[Bates 1978b] R. H. T. Bates. On phase problems: I. *Optik*, 51:161–170, 1978.

[Bates 1978c] R. H. T. Bates. On phase problems: II. *Optik*, 51:223–224, 1978.

[Bates 1982a] R. H. T. Bates. Astronomical speckle imaging. *Physics Reports*, 90:203–297, 1982.

[Bates 1982b] R. H. T. Bates. Fourier phase problems are uniquely solvable in more than one dimension: I - underlying theory. *Optik*, 61:247–262, 1982.

[Bates *et al.* 1984] J. H. T. Bates, W. R. Fright, and R. H. T. Bates. Wiener filtering and cleaning in a general image processing context. *Monthly Notices of the Royal Astronomical Society*, 211, 1984.

[Berenyi *et al.* 1985] H. M. Berenyi, H. V. Deighton, and M. A. Fiddy. The use of bivariate polynomial factorization algorithms in two-dimensional phase problems. *Optica Acta*, 32:689–701, 1985.

[Boas 1954] R. P. Boas. *Entire functions.* Academic Press, New York, 1954.

[Born and Wolf 1970] M. Born and E. Wolf. *Principles of Optics.* Pergammon Press, Oxford, 1970.

[Bracewell 1978] R. N. Bracewell. *The Fourier transform and its applications.* McGraw-Hill, New York, 1978.

[Brames 1986a] B. J. Brames. Sequences with positive semidefinite Fourier transforms. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-34:1502–1510, 1986.

[Brames 1986b] B. J. Brames. Unique phase retrieval with explicit support information. *Optics Letters*, 11:61–63, 1986.

[Bruck and Sodin 1979] Y. M. Bruck and L. G. Sodin. On the ambiguity of the image reconstruction problem. *Optics Communications*, 30:304–308, 1979.

[Bruck and Sodin 1983] Y. M. Bruck and L. G. Sodin. An improved method for reconstruction of two- and multidimensional images from the phase of their Fourier spectrum. *Optica Acta*, 30:995–999, 1983.

[Bruck and Sodin 1984] Y. M. Bruck and L. G. Sodin. Speckle interferometry image reconstruction from the Fourier transform phase. *Journal of the Optical Society of America A*, 1:73–80, 1984.

[Bryan and Skilling 1986] R. K. Bryan and J. Skilling. Maximum entropy image reconstruction from phaseless Fourier data. *Optica Acta*, 33:287–299, 1986.

[Burg 1978a] J. P. Burg. Maximum entropy spectral analysis. In D. G. Childers, editor, *Modern Spectral Analysis*, pages 34–42, IEEE Press, 1978.

[Burg 1978b] J. P. Burg. The relationship between maximum entropy spectra and maximum likelihood spectra. In D. G. Childers, editor, *Modern Spectral Analysis*, pages 132–133, IEEE Press, 1978.

[Burge *et al.* 1976] R. E. Burge, M. A. Fiddy, A. H. Greenaway, and G. Ross. The phase problem. *Proceedings Royal Society of London A*, 350:191–212, 1976.

[Canterakis 1983] N. Canterakis. Magnitude-only reconstruction of two-dimensional sequences with finite regions of support. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-31:1256–1262, 1983.

[Chalasinska-Macukow and Arsenault 1985] K. Chalasinska-Macukow and H. H. Arsenault. Fast iterative solution to exact equations for the two-dimensional phase-retrieval problem. *Journal of the Optical Society of America A*, 2:46–50, 1985.

[Cocke 1985] J. Cocke. The Cauchy-Schwarz inequality as a constraint in power spectrum/autocorrelation analysis and image reconstruction. In *Proceedings SPIE*, 1985.

[Cooley and Tukey 1965] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematical computation*, 19:297–300, 1965.

[Coolidge 1959] J. L. Coolidge. *A treatise on algebraic plane curves.* Constable, London, 1959.

[Cornell 1987] T. J. Cornell. The practice of deconvolution. In *Proceedings of the joint workshop on high-resolution imaging from the ground using interferometric techniques*, pages 177–182, Oracle, Arizona, January 12-15 1987.

[Crimmins and Fienup 1983] T. R. Crimmins and J. R. Fienup. Uniqueness of phase retrieval for functions with sufficiently disconnected support. *Journal of the Optical Society of America*, 73:218–221, 1983.

[Crimmins 1987] T. R. Crimmins. Phase retrieval for discrete functions with support constraints. *Journal of the Optical Society of America A*, 4:124–134, 1987.

[Curtis and Oppenheim 1987] S. R. Curtis and A. V. Oppenheim. Reconstruction of multidimensional signals from zero crossings. *Journal of the Optical Society of America A*, 4:221–231, 1987.

[Curtis *et al.* 1985] S. R. Curtis, A. V. Oppenheim, and J. S. Lim. Signal reconstruction from Fourier transform sign information. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-33:643–657, 1985.

[Dainty and Fienup 1987] J. C. Dainty and J. R. Fienup. Submitted to Image Recovery: Theory and Application, H. Stark Ed., Academic Press, 1987.

[Dainty 1973] J. C. Dainty. Diffraction-limited imaging of stellar objects using telescopes of low optical quality. *Optics Communications*, 7:129–134, 1973.

[Dainty 1982] J. C. Dainty. Stellar interferometry. Available from J. C. Dainty, Blackett Laboratory, Imperial College, London, SW7 2BZ, 1982.

[Davey *et al.* 1986] B. L. K. Davey, A. M. Sinton, and R. H. T. Bates. Zero-and-add. *Optical Engineering*, 25:765–771, 1986.

[Deighton *et al.* 1985] H. V. Deighton, M. S. Scivier, and M. A. Fiddy. Solution of the two-dimensional phase retrieval problem. *Optics Letters*, 10:250–251, 1985.

[Feldkamp and Fienup 1980] G. B. Feldkamp and J. R. Fienup. Noise properties of images reconstructed from fourier modulus. In W. T. Rhodes, editor, *Proceedings SPIE*, pages 84–93, 1980.

[Fiddy *et al.* 1983] M. A. Fiddy, B. J. Brames, and J. C. Dainty. Enforcing irreducibility for phase retrieval in two dimensions. *Optics Letters*, 8:96–98, 1983.

[Fienup and Wackerman 1984] J. R. Fienup and C. C. Wackerman. Improved phase-retrieval algorithm. *Journal of the Optical Society of America A*, 1, 1984.

[Fienup and Wackerman 1986] J. R. Fienup and C. C. Wackerman. Phase retrieval stagnation problems and solutions. *Journal of the Optical Society of America A*, 3:1897–1907, 1986.

[Fienup 1978] J. R. Fienup. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*, 3:27–29, 1978.

[Fienup 1979] J. R. Fienup. Space object imaging through the turbulent atmosphere. *Optical Engnieering*, 18:529–534, 1979.

[Fienup 1982] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21:2758–2769, 1982.

[Fienup 1983a] J. R. Fienup. Comments on: The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-31:738–739, 1983.

[Fienup 1983b] J. R. Fienup. Reconstruction of objects having latent reference points. *Journal of the Optical Society of America*, 73:1421–1426, 1983.

[Fienup 1983c] J. R. Fienup. Reconstruction of objects having latent reference points. *Journal of the Optical Society of America*, 73:1421–1426, 1983.

[Fienup 1987] J. R. Fienup. Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint. *Journal of the Optical Society of America A*, 4:118–223, 1987.

[Fienup *et al.* 1982] J. R. Fienup, T. R. Crimmins, and W. Holsztynski. Reconstruction of the support of an object from the support of its autocorrelation. *Journal of the Optical Society of America*, 72:610–624, 1982.

[Fletcher 1983] R. Fletcher. *Practical methods of optimisation vols 1,2.* John Wiley and Sons, Chichester, 1983.

[Frieden 1985] B. R. Frieden. Dice entropy and likelihood. *Proceedings IEEE*, 73:1764–1770, 1985.

[Fright and Bates 1982] W. R. Fright and R. H. T. Bates. Fourier phase problems are uniquely solvable in more than one dimension: III-computational examples for two-dimensions. *Optik*, 62:333–346, 1982.

[Fright 1984] W. R. Fright. *The Fourier phase problem.* PhD thesis, University of Canterbury, New Zealand, 1984.

[Fright 1987] W. R. Fright. Private communication, 1987.

[Garden 1984] K. L. Garden. *An Overview of Computed Tomography.* PhD thesis, University of Canterbury, New Zealand, 1984.

[Gardenier and Bates 1988] P. H. Gardenier and R. H. T. Bates. Antenna distribution from far field pattern magnitude. In preparation for IEE Proceedings Part H, 1988.

[Gerchberg and Saxton 1972] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[Gerchberg 1986] R. W. Gerchberg. The lock problem in the Gerchberg-Saxton algorithm for phase retrieval. *Optik*, 74:91–93, 1986.

[Goodman 1968] J. W. Goodman. *Fourier Optics.* McGraw-Hill, New York, 1968.

[Gull and Daniell 1978] S. F. Gull and G. J. Daniell. Image reconstruction from incomplete and noisy data. *Nature*, 272:686–690, 1978.

[Haque and Meyer 1986] T. Haque and R. A. Meyer. Iterative reconstruction of images from Fourier transform phase using moments. *Optics Letters*, 11:764–766, 1986.

[Hauptmann 1986] H. Hauptmann. Direct methods and anomalous dispersion. *Agnew. Chem. Int. Ed. Engl.*, 25:603–613, 1986.

[Hayes and McClellan 1982] M. H. Hayes and J. H. McClellan. Reducible polynomials in more than one variable. *Proceedings IEEE*, 70:197–198, 1982.

[Hayes 1982] M. H. Hayes. The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-30:140–154, 1982.

[Hayes *et al.* 1980] M. H. Hayes, J. S. Lim, and A. V. Oppenheim. Signal reconstruction from phase or magnitude. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-28:672–680, 1980.

[Heffernan and Bates 1982] P. B. Heffernan and R. H. T. Bates. Image reconstruction from projections. VI: Comparison of interpolation methods. *Optik*, 60:129–142, 1982.

[Hoenders 1975] B. J. Hoenders. On the solution of the phase retrieval problem. *Journal of Mathematical Physics*, 16:1719–1725, 1975.

[Hofstetter 1964] E. M. Hofstetter. Construction of time limited functions with specified autocorrelation functions. *IEEE Transactions on Information Theory*, IT-10:119–126, 1964.

[Horowitz and Sahni 1976] E. Horowitz and S. Sahni. *Fundamentals of data structures.* Computer Science Press, California, 1976.

[Huang *et al.* 1971] T. S. Huang, W. F. Schreiber, and O. J. Tretiak. Image processing. *Proceedings IEEE*, 59:1586–1609, 1971.

[Izraelevitz and Lim 1987] D. Izraelevitz and J. S. Lim. A new direct algorithm for image reconstruction from Fourier transform magnitude. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-35:511–519, 1987.

[Jaynes 1982] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings IEEE*, 70:939–952, 1982.

[Jenkins and Traub 1972] M. A. Jenkins and J. F. Traub. Algorithm 419 – zeros of a complex polynomial [c2]. *Communications of the ACM*, 15:97–99, 1972.

[Jury 1974] J. I. Jury. *Inners and the stability of dynamic systems.* Wiley, New York, 1974.

[Karle 1986] J. Karle. Recovering phase information from intensity data. *Agnew. Chem. Int. Ed. Engl.*, 25:614–629, 1986.

[Kermisch 1970] D. Kermisch. Image reconstruction from phase information only. *Journal Optical Society of America*, 60:15–17, 1970.

[Kiedron 1981] P. Kiedron. On the 2-D solution ambiguity of the phase recovery problem. *Optik*, 59:303–309, 1981.

[Kreysig 1979] E. Kreysig. *Advanced Engineering Mathematics.* John Wiley and Sons, New York, 1979.

[Labeyrie 1970] A. Labeyrie. Attainment of diffraction-limited resolution in large telescopes by Fourier analysing speckle patterns in star images. *Astronomy and Astrophysics*, 6:85–87, 1970.

[Landau and Pollack 1961] H. J. Landau and H. O. Pollack. Prolate spheroidal wave functions, Fourier analysis and uncertainty-II. *Bell Systems Technical Journal*, 40:65–84, 1961.

[Landau and Pollack 1962] H. J. Landau and H. O. Pollack. Prolate spheroidal wave functions, Fourier analysis and uncertainty-III: the dimension of the space of essentially time- and band-limited signals. *Bell Systems Technical Journal*, 41:1295–1336, 1962.

[Lane and Bates 1987a] R. G. Lane and R. H. T. Bates. Automatic multi-dimensional deconvolution. *Journal of the Optical Society of America A*, 4:180–188, 1987.

[Lane and Bates 1987b] R. G. Lane and R. H. T. Bates. Relevance for blind deconvolution of recovering Fourier magnitude. *Optics Communications*, 63:11–14, 1987.

[Lane 1987] R. G. Lane. Recovery of complex images from Fourier magnitude. *Optics Communications*, 63:6–10, 1987.

[Lane *et al.* 1987] R. G. Lane, W. R. Fright, and R. H. T. Bates. Direct phase retrieval. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-35:520–526, 1987.

[Lawton and Morrison 1987] W. Lawton and J. Morrison. Factoring trigonometric polynomials regarded as entire functions of exponential type. *Journal of the Optical Society of America A*, 4:105–111, 1987.

[Lawton 1980] W. Lawton. A numerical algorithm for 2-d wavefront reconstruction from intensity measurements in a single plane. In *Proceedings of SPIE: International Optical Computing Conference*, pages 94–98, 1980.

[Lawton 1981] W. Lawton. Uniqueness results for the phase retrieval problem for radial functions. *Journal of the Optical Society of America*, 71:1519–1522, 1981.

[Lesem *et al.* 1969] L. B. Lesem, P. M. Hirsh, and J. A. Jordan. The kinoform: a new wavefront reconstruction device. *IBM Journal of Research and Development*, 13:150–155, 1969.

[Levi and Stark 1984] A. Levi and H. Stark. Image restoration by the method of generalised projections with application to restoration from magnitude. *Journal of the Optical Society of America A*, 1:932–943, 1984.

[Levin 1964] B. J. Levin. *Distribution of Zeros of Entire Functions.* American Mathematical Society, Providence, Rhode Island, 1964.

[Makhoul 1975] J. Makhoul. Linear prediction: a tutorial review. *Proceedings IEEE*, 63:561–580, 1975.

[Manolitsakis 1982] I. Manolitsakis. Two-dimensional scattered fields: a description in terms of the zeros of entire functions. *Journal Mathematical Physics*, 26:2141–2298, 1982.

[Markusevich 1965] I. M. Markusevich. *Theory of Functions of a Complex Variable.* Volume 2, Prentice-Hall, Englewood Cliffs, New Jersey, 1965.

[Marr *et al.* 1979] D. Marr, S. Ullman, and T. Poggio. Bandpass channels, zero crossings and early visual information processing. *Journal of the Optical Society of America*, 69:914–916, 1979.

[McCallum 1988] B. McCallum. PhD thesis in preparation, 1988.

[Missell 1978] D. L. Missell. The phase problem in electron microscopy. In V. E. Cosslett and R. Barer, editors, *Advances in Electronics and electron physics*, pages 185–279, Academic Press, 1978.

[Mnyama 1987] D. Mnyama. *Image reconstruction with incomplete data and partial constraints.* PhD thesis, University of Canterbury, New Zealand, 1987.

[Morris 1985] D. Morris. Phase retrieval in the radio holography of reflector antennas and radio telescopes. *IEEE Transactions on Antennas and Propagation*, AP-33:749–755, 1985.

[McDonnell 1975] M. J. McDonnell. *Nonrecursive digital image restoration.* PhD thesis, University of Canterbury, New Zealand, 1975.

[Munson Jr. and Sanz 1986] D. C. Munson, Jr. and J. L. C. Sanz. Phase-only image reconstruction from offset fourier data. *Optical Engineering*, 25:655–661, 1986.

[Nagrath and Gopal 1982] I. J. Nagrath and M. Gopal. *Control Systems Engineering.* Wiley, New York, 1982.

[Nakajima and Asakura 1982] N. Nakajima and T. Asakura. Study of zero location by means of an exponential filter in the phase retrieval problem. *Optik*, 60:289–305, 1982.

[Nakajima and Asakura 1983a] N. Nakajima and T. Asakura. Extraction of the influence of zeros from the image intensity in the phase retrieval using the logarithmic Hilbert transform. *Optik*, 63:99–108, 1983.

[Nakajima and Asakura 1983b] N. Nakajima and T. Asakura. Phase retrieval from the image intensity using an exponential filter with the purpose of reducing the influence of zeros. *Optik*, 64:37–49, 1983.

[Nakajima and Asakura 1985] N. Nakajima and T. Asakura. A new approach to two-dimensional phase retrieval. *Optica Acta*, 32:647–658, 1985.

[Nakajima 1986] N. Nakajima. Improvement in evaluating the logarithmic Hilbert transform in phase retrieval. *Optics Letters*, 11:600–602, 1986.

[Napier and Bates 1974] P. J. Napier and R. H. T. Bates. Inferring phase information from modulus information in two-dimensional aperture synthesis. *Astronomy and Astrophysics Supplement Series*, 15:427–430, 1974.

[Napier 1971] P. J. Napier. *Reconstruction of Radiating Sources.* PhD thesis, University of Canterbury, New Zealand, 1971.

[Narayan 1987] N. Narayan. Phase retrieval with the maximum entropy method. In *Proceedings of the joint workshop on high-resolution imaging from the ground using interferometric techniques*, pages 183–186, Oracle, Arizona, January 12-15 1987.

[Nieto-Vesperinas and Dainty 1984] M. Nieto-Vesperinas and J. C. Dainty. Testing for uniqueness of phase recovery in two-dimensions. *Optics Communications*, 52:94–98, 1984.

[Nieto-Vesperinas and Dainty 1985] M. Nieto-Vesperinas and J. C. Dainty. A note on Eisenstein's irreducibility ctiterion for two-dimensional sampled objects. *Optics Communications*, 54:333–334, 1985.

[Nieto-Vesperinas and Dainty 1986] M. Nieto-Vesperinas and J. C. Dainty. Phase recovery for two-dimensional digital objects by polynomial factorisation. *Optics Communications*, 58:83–88, 1986.

[Nieto-Vesperinas and Mendez 1986] M. Nieto-Vesperinas and J. A. Mendez. Phase retrieval by Monte-Carlo methods. *Optics Communications*, 59:249–254, 1986.

[Nityananda and Narayan 1982] R. Nityananda and R. Narayan. Maximum entropy image reconstruction - a practical non-information-theoretic approach. *Journal Astrophysics and Astronomy*, 3:419–450, 1982.

[Nowinski 1981] J. L. Nowinski. *Applications of functional analysis in engineering.* Plenum Press, New York and London, 1981.

[Nussensveig 1972] H. M. Nussensveig. *Causality and Dispersion Relations.* Academic Press, New York and London, 1972.

[Oppenheim and Lim 1981] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. *Proceedings IEEE*, 69:529–541, 1981.

[Oppenheim and Schafer 1975] A. V. Oppenheim and R. W. Schafer. *Digital Signal processing.* Prentice-Hall, New Jersey, 1975.

[Oppenheim *et al.* 1968] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr. Nonlinear filtering of multiplied and convolved signals. *Proceedings IEEE*, 56:1264–1291, 1968.

[Oppenheim *et al.* 1982] A. V. Oppenheim, M. H. Hayes, and J. S. Lim. Iterative procedures for signal reconstruction from Fourier transform phase. *Optical Engineering*, 21:122–127, 1982.

[Papoulis 1984] A. Papoulis. *Signal analysis.* McGraw-Hill, London, 1984.

[Rabiner and Gold 1975] L. R. Rabiner and B. Gold. *Theory and Applications of Digital Signal Processing.* Prentice-Hall, New Jersey, 1975.

[Ramachandran and Srinivasan 1970] G. N. Ramachandran and R. Srinivasan. *Fourier Methods in Crystallography.* Wiley-Interscience, New York, 1970.

[Readhead *et al.* 1980] A. C. S. Readhead, R. C. Walker, T. J. Pearson, and C. M. H. . Mapping radio sources with uncalibrated visibility data. *Nature*, 285:137–140, 1980.

[Requicha 1980] A. A. G. Requicha. The zeros of entire functions: theory and engineering applications. *Proceedings IEEE*, 68:308–328, 1980.

[Roddier 1981] F. Roddier. The effects of atmospheric turbulence in optical astronomy. In E. Wolf, editor, *Progress in Optics*, pages 283–376, North-Holland, 1981.

[Ross *et al.* 1978] G. Ross, M. A. Fiddy, M. Nieto-Vesperinas, and M. W. L. Wheeler. The phase problem in scattering phenomena: the zeros of entire functions and their significance. *Proceedings Royal Society of London A*, 360:25–45, 1978.

[Rotem and Zeevi 1986] D. Rotem and Y. Y. Zeevi. Image reconstruction from zero crossings. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-34:1269–1277, 1986.

[Sanz and Huang 1983a] J. L. C. Sanz and T. S. Huang. Unified Hilbert space approach to iterative least-squares linear signal restoration. *Journal of the Optical Society of America*, 73:1455–1465, 1983.

[Sanz and Huang 1983b] J. L. C. Sanz and T. S. Huang. Unique reconstruction of a band-limited multidimensional signal from its phase or magnitude. *Journal of the Optical Society of America*, 73:1446–1450, 1983.

[Sanz and Huang 1985] J. L. C. Sanz and T. S. Huang. Polynomial system of equations and its applications to the study of the effect of noise on multidimensional Fourier transform phase retrieval from magnitude. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-33:997–1004, 1985.

[Sanz 1985a] J. L. C. Sanz. Mathematical considerations for the problem of Fourier transform phase retrieval from magnitude. *SIAM Journal of Applied Mathematics*, 45:651–664, 1985.

[Sanz 1985b] J. L. C. Sanz. On the reconstruction of band-limited multidimensional signals from algebraic sampling contours. *Proceedings IEEE*, 73:1334–1336, 1985.

[Sanz *et al.* 1983] J. L. C. Sanz, T. S. Huang, and F. Cukierman. Stability of unique Fourier-transform phase reconstruction. *Journal of the Optical Society of America*, Journal of the Optical Society of America:1442–1445, 1983.

[Sanz *et al.* 1984] J. L. C. Sanz, T. S. Huang, and T. F. Wu. A note on iterative Fourier transform phase reconstruction from magnitude. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-32:1251–1254, 1984.

[Sasaki and Yamagami 1987] O. Sasaki and T. Yamagami. Phase-retrieval algorithms for nonnegative and finite-extent objects. *Journal of the Optical Society of America A*, 4:720–726, 1987.

[Saxton 1974] W. O. Saxton. Phase determination in bright-field electron microscopy using complementary half-plane apertures. *Journal Physics D: Applied Physics*, 7:L63–L64, 1974.

[Sayegh *et al.* 1987] S. I. Sayegh, Y. L. Kok, and J. H. Hong. An algorithm to find two-dimensional signals with specified zero crossings. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-35:107–111, 1987.

[Schafer *et al.* 1981] R. W. Schafer, R. M. Mersereau, and M. A. Richards. Constrained iterative algorithms. *Proceedings IEEE*, 69:432–450, 1981.

[Scivier and Fiddy 1985a] M. S. Scivier and M. A. Fiddy. Ambiguities in magnitude-only reconstruction of band-limited signals. *Optics Letters*, 10:369–371, 1985.

[Scivier and Fiddy 1985b] M. S. Scivier and M. A. Fiddy. Phase ambiguities and the zeros of multidimensional band-limited functions. *Journal of the Optical Society of America A*, 2:693–697, 1985.

[Sezan and Stark 1982] M. I. Sezan and H. Stark. Image restoration by the method of convex projections: Part 2 - Applications and numerical results. *IEEE Transactions on Medical Imaging*, MI-1:95–101, 1982.

[Shannon 1948] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[Silver 1965] S. Silver, editor. *Microwave Antenna Theory and Design.* Dover Publications, New York, 1965.

[Sinton 1986] A. M. Sinton. *Contributions to Astronomical and Medical Information Processing.* PhD thesis, University of Canterbury, New Zealand, May 1986.

[Sinton *et al.* 1986] A. M. Sinton, B. L. K. Davey, and R. H. T. Bates. Augmenting shift-and-add with zero-and-add. *Journal of the Optical Society of America A*, 3:1010–1017, 1986.

[Slepian and Pollack 1961] D. Slepian and H. O. Pollack. Prolate spheroidal wave functions, Fourier analysis and uncertainty-I. *Bell Systems Technical Journal*, 40:43–63, 1961.

[Slepian 1976] D. Slepian. On bandwidth. *Proceedings IEEE*, 64:292–300, 1976.

[Slepian 1978] D. Slepian. Fourier analysis and uncertainty V: the discrete case. *Bell Systems Technical Journal*, 57:1371–1430, 1978.

[Slepian 1983] D. Slepian. Some comments on Fourier analysis, uncertainty and modelling. *SIAM Review*, 1983:379–394, 1983.

[Stefanescu 1985] I. S. Stefanescu. On the phase retrieval problem in two dimensions. *Journal Mathematical Physics*, 26:2141–2160, 1985.

[Stockham Jr. *et al.* 1975] T. G. Stockham, Jr., T. M. Cannon, and R. B. Ingebretson. Blind deconvolution through digital signal processing. *Proceedings IEEE*, 63:678–692, 1975.

[Taylor and Whinnery 1951] T. T. Taylor and J. R. Whinnery. Applications of potential theory to the design of linear arrays. *Journal Applied Physics*, 22:19–29, 1951.

[Taylor 1982] J. R. Taylor. *An introduction to error analysis.* University Science, Mill Valley, California, 1982.

[Titchmarsh 1932] E. C. Titchmarsh. *The Theory of Functions.* Clarendon Press, Oxford, 1932.

[Tom *et al.* 1981] V. T. Tom, T. F. Quatieri, M. H. Hayes, and J. H. McClellan. Convergence of non-expansive signal reconstruction algorithms. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-29, 1981.

[Toraldo di Francia 1955] G. Toraldo di Francia. Resolving power and information. *Journal of the Optical Society of America*, 45:497–501, 1955.

[Tribolet 1977] J. M. Tribolet. A new phase unwrapping algorithm. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-25:170–177, 1977.

[Trussell and Civanlar 1984] H. J. Trussell and M. R. Civanlar. Feasible solutions in signal restoration. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-32:201–212, 1984.

[van den Bos 1971] A. van den Bos. Alternative interpretation of maximum entropy spectral analysis. *IEEE Transactions on Information Theory*, IT-17:493–494, 1971.

[van Toorn and Ferweda 1977] P. van Toorn and H. A. Ferweda. On the phase problem retrieval in electron microscopy from image and diffraction patterns, IV - checking of algorithm by means of simulated objects. *Optik*, 47:123–134, 1977.

[Walker 1950] R. J. Walker. *Algebraic curves.* Princeton University Press, Princeton, New Jersey, 1950.

[Walker 1981a] J. G. Walker. Object reconstruction from turbulence degraded images. *Optica Acta*, 28:1017–1019, 1981.

[Walker 1981b] J. G. Walker. The phase retrieval problem. A solution based on zero location by experimental apodisation. *Optica Acta*, 28:735–738, 1981.

[Walther 1963] A. Walther. The question of phase retrieval in optics. *Optica Acta*, 11:41–49, 1963.

[Wang and Lim 1982] D. L. Wang and J. S. Lim. The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-30:679–681, 1982.

[Wiener 1942] N. Wiener. *Extrapolation, Interpolation and smoothing of stationary time series.* MIT Press, Cambridge, Massachusetts, 1942.

[Wilkinson 1963] J. H. Wilkinson. *Rounding errors in algebraic processes.* Her Majesty's Stationary Office, London, 1963.

[Won *et al.* 1985] M. C. Won, D. Mnyama, and R. H. T. Bates. Improving initial phase estimates. *Optica Acta*, 32:377–396, 1985.

[Woodward 1964] P. M. Woodward. *Probability and Information Theory with Applications to Radar.* Pergamon Press, Oxford, 1964.

[Yeh and Chin 1985] C. L. Yeh and R. T. Chin. Error analysis of a class of constrained iterative algorithms. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-33:1593–1598, 1985.

[Youla and Webb 1982] D. C. Youla and H. Webb. Image restoration by the method of convex projections: Part 1 - Theory. *IEEE Transactions on Medical Imaging*, MI-1:81–94, 1982.

[Youla 1978] D. C. Youla. Generalised image restoration by method of alternating projections with application to restoration from magnitude. *IEEE Transactions on Circuits and Systems*, CAS-25:694–702, 1978.

[Zakhor and Izraelevitz 1986] I. Zakhor and D. Izraelevitz. A note on the sampling of zero crossings of two-dimensional signals. *Proceedings IEEE*, 74:1285–1287, 1986.